PERSONALIZED CAMPAIGN RECOMMENDATION AND BUYER TARGETING FOR B2B MARKETING

by

JINGYUAN YANG

A dissertation submitted to the

Graduate School-Newark

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Management

written under the direction of

Dr. Hui Xiong

and approved by

Newark, New Jersey

May 2018

ABSTRACT OF THE DISSERTATION

Personalized Campaign Recommendation and Buyer Targeting for B2B Marketing

By JINGYUAN YANG

Dissertation Director: Dr. Hui Xiong

Business to Business (B2B) marketing aims at meeting the needs of other businesses instead of individual consumers, and thus entail management of more complex business needs than consumer marketing. The buying processes of the business customers involve series of different marketing campaigns providing multifaceted information about the products or services. While most existing studies focus on individual consumers, little has been done to guide business customers due to the dynamic and complex nature of these business buying processes. In this dissertation, we focus on providing data-driven solutions to achieve two important business goals: reduce the buying cycle time and increase the conversion rate.

Specifically, we first introduce a unified view of social and temporal modeling for B2B marketing campaign recommendation to reduce the buying cycle time. Along this line, we exploit the temporal behavior patterns in the buying processes of the business customers and develop a B2B marketing campaign recommender system. Specifically, we first propose the temporal graph as the temporal knowledge representation of the buying process of each business customer. We then develop the low-rank graph reconstruction framework to identify the common graph patterns and predict the missing edges in the temporal graphs. In addition, we also exploit the community relationships of the business customers to improve the performances of the

graph edge predictions and the marketing campaign recommendations. Results from extensive empirical studies on real-world B2B marketing data sets show that the proposed method can effectively improve the quality of the campaign recommendations for challenging B2B marketing tasks.

Furthermore, we develop two different approaches aiming at improving the conversion rate. We first present a novel unified framework to integrate two important marketing tasks–customer segmentation and buyer targeting. Instead of combining these two tasks in a simple step-by-step approach, we formulate customer segmentation and buyer targeting as a unified optimization problem. Then, the customer segments are adaptively realized during the targeting optimization process. In this way, the integrated approach not only improves the buyer targeting performances but also provides a new perspective of segmentation based on the buying decision preferences of the customers. Finally, we introduce a predictive lead scoring model which can help sales representatives to identify prospective leads from a large pool of candidates in a B2B environment. Specifically, we provide a multi-focal lead scoring framework which can improve the performance of predictive lead scores. However, independent modeling at focal level would be problematic for segments with few representative samples. We use the Multi-Task Learning framework to address this problem by exploiting commonalities shared by focal groups and automatically balancing between unification of all groups and individualization of each group. Therefore, such a multi-focal tailored lead scoring model gives a better insight into factors influencing the conversion of leads.

ACKNOWLEDGEMENTS

It is a sentimental moment to write this note of thanks as the finishing touch of my dissertation. This thesis represents the cohesion of painstaking efforts of several years of work, and could not have been accomplished without the love and support from many people. Here I would like to express my greatest appreciation to all the people who provided me tremendous support and help during my Ph.D. study.

I would like first to express my deep gratitude to my mentor and adviser, Professor Hui Xiong.He provide me intellectual guidance, inspiring insights, inexhaustible patience, and extraordinary knowledge, which are necessary to survive and thrive the graduate school and the beyond. I thank him for generously giving me motivation, support, time, understandings, assistance, opportunities, trust, and friendship for all these years. Moreover, I want to thank him for providently positioning and guiding me on the right direction of the pioneer research work of data mining and marketing analytics. He helped me to be a better writer, speaker, and scholar.

I also sincerely grateful to those who served in my committees: Prof. Xiaodong Lin, Prof. Spiros Papadimitriou, Prof. Dacheng Tao and Prof. Jian Yang. I deeply appreciate their time and efforts in reviewing my dissertation and attending my dissertation defense. All of them not only provide thought-provoking and constructive suggestions on my research, but also offer tremendous support and help in my academic job hunting process. Special thanks are due to Prof. Yong Ge at Uniersity of Arizona, Prof. Meng Qu at Rutgers University, Prof. Keli Xiao at SUNY-Stony Brook, Prof. Wenjun Zhou at University of Tennessee, Prof. Chuanren Liu at Drexel University, Prof. Yu Du at University of Colorado Denver, and Prof. Yanjie Fu at Missouri University of Science and Technology and for their help with both my job search and research work. I also owe a hefty amount of thanks to the faculty and student members in the Data Mining group of Rutgers Business School: Dr. Hengshu Zhu, Dr. Yanchi Liu, Dr. Bin Liu, Prof. Fuzhen Zhuang, Mingfei Teng, Huang Xu, Zijun Yao, Junming Liu, Qiao Li, Konstantin Patev, Farid Razzak, Hao Zhong, Dr. Zhongmou Li, Can Chen, Qingxin Meng for their help, friendship, and valuable suggestions. It is an honor to be affiliated with such a great group.

I would also like to acknowledge the Department of Management Science and Information Systems (MSIS) for supplying me with the best equipment and working environment that helped me to accomplish much of this work. I want to particularly thank Mrs. Luz Kosar and Mr. Goncalo Filipe for all of their assistance.

Finally, my deepest thanks and love to my husband Xiang Liao, and my sweet daughter Erica, who gave me infinite joy, love and support to make this work possible. I owe my deepest gratitude and love to my parents and in-laws for their unwearied devotion and unconditional love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Business-to-Business (B2B) marketing involves marketing of one company product or service to another company, the business customers. In the big data era, massive amounts of marketing data, such as demographics, firmographics, and customer behaviors, accumulated in marketing industry has greatly changed the paradigm of marketing services. These valuable marketing data are fine-grained and informative, and thus provide exceptional opportunities for marketing professionals to discover and understand the relationship among customer behaviors and intentions, which can ultimately assist intelligent business decision-making. Data driven marketing analytics has been utilized as a powerful tool to process marketing data for achieving the final goal of increasing and expediting conversions and boosting profits. Indeed, the focus of this dissertation is to develop efficient and effective techniques to address various B2B marketing challenges by addressing the unique characteristics of real world B2B marketing data with temporal dynamics and unprecedented complexity.

In this chapter, we provide an introduction of B2B marketing analytics. We first provide an introduction of the background and preliminaries, briefly discuss the B2B marketing data set, summarize the contributions of this collection of research, and overview the major contents of this thesis.

1.1 Background & Preliminaries

In this section, we provide a brief overview of the context and preliminaries in B2B marketing field.

1.1.1 Business-to-Business (B2B) Marketing

In general, marketing serves for sales to better sell products or provide services to the customers. During marketing process, the business will let the customers interact with a variety of marketing campaigns/or events to make the customers be better aware their products or service and finally make the buying decision.

	B2B	**B2C**
Buyer	Multiple	Single
Buying Process	Multiple Steps	Single Step
Sales Cycle	Long	Short
Money Volume	High	Low

Table 1.1: Comparison between B2B and B2C marketing.

B2B vs. B2C marketing. Due to the differences between B2B and B2C (Business-to-Consumer), the marketing processes for them are also different in four aspects as shown in the Table 1.1:

1. Buyer: The buyer or customer for B2B is often a group or committee comprising

multiple individuals who have different roles in evaluating the product/services. While in B2C, the buyer is just a single person who needs the product/service for personal use.

2. Buying Process: A standard B2B buying process usually consists of multiple steps, which represent progressive stages a customer goes through before making final buying decision. But in B2C the buying process is a single step.

3. Sales Cycle: The sale cycle of B2B is thus much longer than B2C sales cycle due to more complex buying process.

4. Money Volume: B2B transactions are significantly higher in money value than B2C transactions. The B2B market is 5 times as large as the B2C market in the U.S.

Important Business Goals. Along the B2B marketing process, there are two important business goals that marketers want to achieve:

- **Increase The Conversion Rate**: In general, the B2B marketing process is just like a funnel process, although there are a large number of visitors/contacts at beginning stage, only a small number of potential visitors are converted into final paying customers. In the real-world, the average conversion rate is usually as low as around 2-3%. Thus, to achieve a higher conversion rate is one of the top business goal for B2B marketing.

- **Reduce The Cycle Time**: The B2B marketing business process is usually very long and complex. The marketing professionals are dedicated to reduce the cycle time to accelerate the buying process and increase the conversion velocity.

To achieve these two crucial business goals, novel approaches to improve the effectiveness and efficiency of the marketing strategies will be needed. Advanced marketing analytics is promised to provide marketers more intelligent and customized solutions by applying data mining techniques to understand and investigate the large-scale marketing data set.

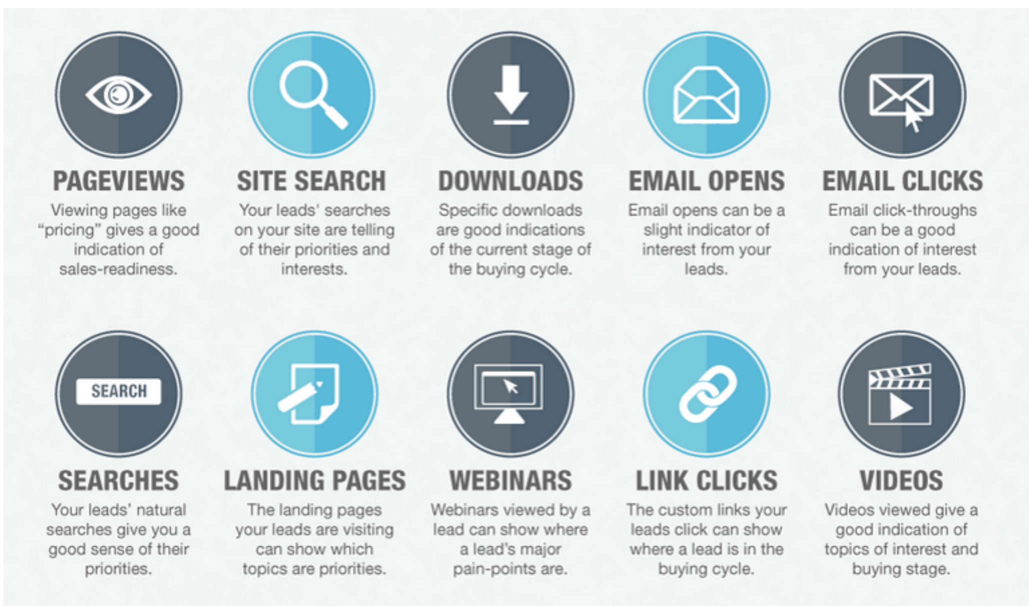


Figure 1.1: Examples of customer interaction behavior.

B2B Marketing Data. Commonly, there are two types of marketing data: 1) **Static** data includes *demographic* information about each customers and *firmographic* information describing the organizational information, such as industry, company size, years in business, and so on; 2) **Dynamic** customer behavior data that describe how these customers interact with different marketing campaigns through the buying process. Figure 1.1 shows some examples of typical customer interactions, such as downloading product trial, attending webinars, opening newsletter email, etc.

By utilizing the combination of information-rich static and dynamic marketing data, we can better understand real buying intention and readiness behind the customers' marketing campaign behavior, which in turn helps us quickly increase the conversion rate. Moreover, once we identify the customers in which specific buying stage, we are able to provide proper marketing campaign recommendation, which can help the customer move to the next level of the purchasing cycle, ultimately helping reduce the cycle time as well.

1.1.2 Research Motivation

While most existing studies focus on individual consumers, little has been done to guide business customers due to the dynamic and complex nature of these business buying processes. The buying processes in B2B markets usually involve series of different marketing campaigns providing multi-faceted information to multiple decision makers with different focal points and motivations. Related to the characteristics of B2B marketing as described in Table 1.1, there are several unique challenges we are facing when developing B2B marketing solutions. First, varies marketing campaigns are developed to improve the low conversion rate, however offering and managing campaigns at such volume on such a large customer base poses a significant labor and financial burden on the marketers. In practice, most of marketing strategies are made by marketers based on experience. However, as business grows, number of marketing campaigns as well as business customers could increase to a daunting level for humans to handle. There is a critical need for an automated process to provide guidance and improve efficiency. Second, the temporal information is critical to understand the customers behavior pattern. The B2B buying process is very long and involves a number of different marketing campaigns selected from a large pool of campaigns. Potential buyers interact with different campaigns based on their current status to get different facets of information to make further decisions. Therefore the temporal information, especially the temporal correlation/ dependency among the marketing campaigns, is very useful for understanding the needs of the customers and modeling the customer behaviors, specifically providing crucial insights about not only the preferences of the campaign types, but also the favorable campaign orders for execution. However, the implicit temporal information is hidden in the noisy buying processes and we need robust statistics to capture the temporal dynamics. Third, there are usually multiple decision makers from a company evaluating the potential products or services from different aspects. Accordingly, the individuals/customers from the same company can form a committee or community, where each individual can have impacts on the behavior preferences of others in the same community. Thus, to provide tailored marketing strategies to potentially expedite the overall conversion cycle, we need to consider not only each individual's own status, but also the overall behavior preferences of his/her colleagues in the same community.

To this end, we propose three novel marketing techniques aiming at different perspectives but with consideration of the aforementioned aspects, namely, 1) a B2B marketing campaign recommender system with a unified view of social and temporal modeling, and 2) a predictive buyer targeting model with a unified customer segmentation perspective; 3) a multi-focal lead scoring framework for predictive modeling in B2B market.

1.2 Research Contributions

In this dissertation, we study the unique characteristics of B2B marketing data and demonstrate how to develop data-driven solutions with different focuses. Generally, the proposed collection of research has the following major contributions:

- Investigating the impact of the unique characteristics of B2B customer campaign behavior data on the development of campaign recommender systems to reduce the buying cycle time. To this end, we will exploit customer behavior data and form the campaign selection problem into a "Next Campaign To Run" (NCTR) recommender system task, and then present two approaches to develop an innovative campaign recommender system with an integration of temporal and social factors. In addition, the campaign recommder system will also have the backward compatibility to model customer's latent interest as in conventional movie/product recommender systems.

- Development of a novel approach to integrate two important marketing tasks to improve the conversion rate. The key idea is to unify customer segmentation and buyer targeting tasks into an optimization framework. In this way, instead of independently performing the two tasks in a step-by-step approach, we can jointly implement these two tasks in a more integrated and optimized way. Thus, the integrated approach not only improves the buyer targeting performances but also provides a new perspective of segmentation based on the buying decision preferences of the customers.

- Design of a multi-focal lead scoring framework that can significantly improve the quality of lead prospecting for B2B marketing applications. In particular, this multi-focal framework allows marketers to gain better marketing insights, such as what types of leads or lead characteristics matter most. Our findings provide empirical evidence about how leads interact with marketing nurturing campaigns and shed new light on the drivers of lead responses to a firm's marketing campaigns. Moreover, the proposed framework has a potential to be generalized to other marketing and business scenarios, such as brand promotion, product cross-selling or up-selling, and targeted advertising.

Specifically, we first provide a novel B2B marketing campaign recommender system to predict the customized "next-campaign" for reducing the cycle time and expediting the conversion process. Concretely, a personalized temporal graph for each business customer is constructed based on the customer's campaign participating sequence to extract and integrate the campaign order preferences as the temporal knowledge representation of the buying process. The campaign recommender is then built in a low-rank graph reconstruction framework to identify the common graph patterns and predict missing edges in the temporal graphs. The prediction of the unobserved graph edges is effective to recommend the marketing campaigns to the business customers during their buying processes. Also, we incorporate social factors, such as community relationships of the business customers, for further improving overall performances of the missing edge prediction and recommendation. Moreover, we further improve the NCTR framework with the backward compatibility and present the low-rank temporal graph reconstruction as a probabilistic graphical model. The graphical model intuitively demonstrates the components (and their statistical relationship) in the graph reconstruction framework and also incorporates appropriate distribution priors to avoid over-fitting issues. Extensive empirical studies on real-world B2B marketing data sets and the results show that the proposed method can effectively improve the quality of the campaign recommendations for challenging B2B marketing tasks.

Second we provide another focused study of a buyer targeting optimization framework aiming at improving the conversion rate and providing interpretable segmentation simultaneously. More specifically, we first provide a novel mathematical formulation to integrate the customer segmentation and the buyer targeting into a unified optimization problem. Then, we develop an iterative algorithm (*K-Classifiers Segmentation*) to optimize the customer segmentation and targeting simultaneously. The customers assigned to the same targeting model form one customer segment, where the customers buying decision preferences can be explained by the associated targeting model. As a result, targeting performance can be improved due to the buying decision oriented segmentation. To improve the interpretability and the robustness of the results, we further develop a *profile-consistent K-Classifiers Segmentation* algorithm. With the *profile-consistent* algorithm, the identified segmentation is consistent with not only the customer profiles but also the customer decision preferences. Empirical studies on both synthetic and real-world data sets show promising results on not only increased targeting accuracy but also meaningful segmentation with actionable marketing implications.

Third we demonstrate a predictive lead scoring model which can help sales representatives to identify prospective leads from a large pool of candidates in a B2B environment. Specifically, we provide a multi-focal lead scoring framework which can improve the performance of predictive lead scores by exploring discrepancy among lead segments. In this framework, leads are first divided into several focal groups (segments) based on their characteristic attributes (features) and marketing workflows. Then, a logistic regression scoring model is learned for each segment with multi-task learning (MTL) technique. Indeed, the key of multi-focal learning in this study is to allow predictive modeling in each segment consisting of leads with similar characteristics rather than modeling the whole population of leads with varying characteristics. However, independent modeling at focal level would be problematic for segments with few representative samples. We use the MTL framework to address this problem by exploiting commonalities shared by focal groups and automatically balancing between unification of all groups and individualization of each group. Finally, empirical findings derived from real-world B2B marketing data demonstrate that different segments may have absolutely different conversion rates and leads in the same segment tend to have similar responses to a specific marketing campaign.

1.3 Overview

Chapter 2 presents different types of low-rank temporal graph reconstruction models for B2B marketing campaign recommendation. Two different ways are introduced to represent the temporal knowledge of each customer's buying process. *NCTRC* is the low-rank temporal graph reconstruction approach with the community regularization for customers from the same company. Then, *NCTRC* is extended into *NCTRG* with separately defined temporal preferences and campaign interestingness and regularization terms in a graphical model to avoid overfitting. Experimental results and a case study with real world data are presented to validate the effectiveness of campaign recommendations.

Chapter 3 presents a unified optimization framework for buyer targeting. First, we innovatively integrate customer segmentation and buyer targeting in a unified optimization framework. Then, a *K-Classifiers Segmentation* algorithm and the *profile consistent* extension are proposed to optimize the customer segmentation and targeting simultaneously. Finally, we validate our approach on both a synthetic and real-world customer data sets. Experimental results clearly show the effectiveness of our approach and theinterpretable customer segmentation solutions reveals new marketing insights.

Chapter 4 discuss a multi-focal lead scoring framework for improving performance of predictive modeling in B2B market. Our multi-focal lead scoring framework consists of two phases. The first phase is to form focal groups, where each focal group is one unique segment of leads on market. In the second phase, we jointly build lead scoring models for multiple focal groups of customers by using Multi-Task Learning (MTL). The experimental studies show that our approach leads to better learning performances than conventional lead scoring methods.

CHAPTER 2

B2B MARKETING CAMPAIGN RECOMMENDER SYSTEM

Business to Business (B2B) marketing aims at meeting the needs of other businesses instead of individual consumers, and thus entails management of more complex business needs than consumer marketing. The buying processes of the business customers involve series of different marketing campaigns providing multifaceted information about the products or services. While most existing studies focus on individual consumers, little has been done to guide business customers due to the dynamic and complex nature of these business buying processes. To this end, in this chapter, we focus on providing a unified view of social and temporal modeling for B2B marketing campaign recommendation. Along this line, we first exploit the temporal behavior patterns in the B2B buying processes and develop a marketing campaign recommender system. Specifically, we start with constructing temporal graph as the knowledge representation of the buying process of each business customer. Temporal graph can effectively extract and integrate the campaign order preferences of individual business customers. It is also worth noting that our system is backward compatible since the participating frequency used in conventional static recommender systems is naturally embedded in our temporal graph. The campaign recommender is then built in a low-rank graph reconstruction framework based on probabilistic graphical models. Our framework can identify the common graph patterns and predict missing edges in the temporal graphs. In addition, since business customers very often have different decision makers from the same company, we also incorporate social factors, such as community relationships of the business customers, for further improving overall performances of the missing edge prediction and recommendation. Finally, we have performed extensive empirical studies on real-world B2B marketing data sets and the results show that the proposed method can effectively improve the quality of the campaign recommendations for challenging B2B marketing tasks.

2.1 Introduction

Business-to-Business (B2B) marketing involves marketing of one company product or service to another company, the business customers. The B2B market is the largest of all the markets, and exceeds the consumer market in dollar value. These marketing campaigns help in every layer of the sales funnel, all the way to the final goal of increasing and expediting conversions and boosting profits. Success of the B2B marketing always lies in choosing the right campaigns and the right timing. Today most of such decisions are made by marketers based on experience. However, as business grows, number of marketing campaigns as well as business customers could increase to a daunting level for humans to handle. For example, in our collected B2B marketing data from a Fortune 500 software company, one business marketer offered 24,125 marketing campaigns to 2,119 customers, each with a unique series of campaigns, from January 2013 to December 2014. Offering and managing campaigns at such volume on such a large customer base poses a significant labor and financial burden on the marketers. There is a critical need for an automated process to provide guidance and improve efficiency. Therefore, we propose a B2B marketing campaign recommender system, which can not only reduce the burden and cut costs, but also better serve customer needs and lead to better marketing performance.

Developing such a B2B marketing campaign recommender system, however, is a nontrivial task. The buying processes in B2B markets usually involve series of different marketing campaigns providing multi-faceted information to multiple decision makers with different focal points and motivations. These processes are naturally dynamic and complex. As a result, to recommend the right campaign at the right time, it is important to identify the customer behavior patterns hidden in the buying processes, so as to meet the dynamically changing information needs of the business customers. More specifically, we summarize the unique challenges for providing recommendation for "Next Campaign To Run" (**NCTR**) in B2B marketing as follows:

- **Temporal Dynamics**. During the different stages of the buying process, the business customers often engage in different campaigns for information most relevant to them at that particular point of time. This information is then necessary for them to make further decisions and possible move on to the next stage. Therefore the temporal information, especially the temporal correlation/dependency among the marketing campaigns, is very useful for understanding the needs of the customers and modeling the customer behaviors. In other words, temporal-aware models are essential for providing effective marketing campaign recommendations. However, the implicit temporal information is hidden in the noisy buying processes and we need robust statistics to capture the temporal dynamics.

- **Backward Compatibility**. By backward compatibility, we mean that the 'latent interest' of each decision maker should still be modeled as in the conventional static recommender system. In conventional movie/product recommender systems, the interaction frequency between the user and item (e.g., movie, product) is often used as the interest indicator [50, 40]. We would like to keep this backward compatibility in our B2B marketing campaign recommender system. Therefore, the campaign participating frequency of the business customers should also be modeled, together with temporal dynamics in the buying process.

- **Multiple Decision-makers**. For B2B marketing, there are usually multiple decision makers from the same business/company evaluating the potential products or services from different aspects. Accordingly, the individuals/customers from the same company can form a community, where each individual can have impacts on the behavior preferences of others in the same community. Thus, to provide recommendations potentially expediting the overall conversion cycle, we need to consider not only each individual's own status, but also the overall behavior preferences of his/her colleagues in the same community.

In the literature, there have been some related work on the "next-item" recommendation making attempts at some of these challenges we are facing, though mostly from B2C marketing perspective. For example, a factorization framework of Markov chains (FPMC) is proposed by [57, 17]. The idea is to transform the sequential data of each user into a transition matrix and then predict the users next action by factorizing the matrices. Yap et al. [74] search the personalized sequential patterns for next-item recommendation. In addition, Zhao et al. [77, 78] consider the temporal intervals between the purchase behaviors to increase the temporal diversity in the recommendations. Generally speaking, these approaches first extract the temporal knowledge to capture the temporal dynamics of user's preference and then integrate the knowledge in the recommender system. Although these methods have been successfully applied in the B2C (Business-to-Consumer) markets, they are not designed for our B2B marketing scenarios:

- First, the temporal dynamics of individual consumer's preferences on different products are different from the business customers' need on B2B campaigns. In particular, in B2B buying process, different campaigns reveal different levels/facets of information about the same buying process, while such evolution of levels/facets is rare in campaigns for consumer products.

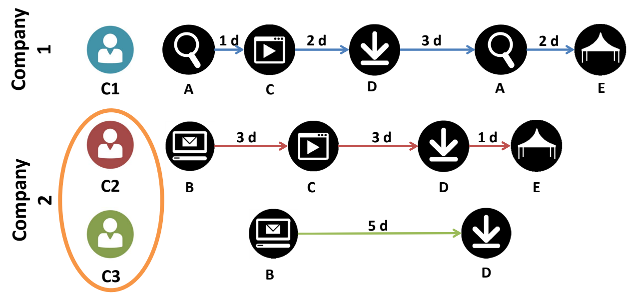
- Second, the existing temporal knowledge representations, such as Markov transition matrix and personalized sequential patterns, are not resistant to noisy behavior records in the complex buying processes. For example, the business customer may occasionally participate in campaigns irrelevant to the current context in the process. This kind of random behavior is sometimes strategic and necessary in B2B marketing but may dramatically affect the recommendation system if we adopt simple Markov transitions or sequential patterns as frequently used in B2C recommendation.

- Third, the multiple decision-makers scenario in B2B marketing makes community structure crucial to recommending a holistic solution for the company as a whole. However, only a scarce of studies have integrated both the individual behavior records and the community structures within a unified recommender system. Our integration of the community structures in the B2B campaign recommender system is motivated by the work of [43, 53, 75].

To address all the challenges and solve the Next Campaign To Run (NCTR) problem for B2B marketing campaign recommendations, we present a novel *NCTR* campaign recommender system. Specifically, we start with the adoption of temporal graph as the temporal knowledge representation of the buying process of each business customer. The key idea is to extract and integrate the campaign order preferences of the customer using the temporal graph. Next, we develop the low-rank graph reconstruction framework to identify the common graph patterns and predict the unobserved edges in the temporal graphs. We showed that the prediction of the unobserved graph edges is effective to recommend the marketing campaigns to the business customers during their buying processes. In addition, to exploit the community structure of the business customers for marketing campaign recommendation, we formulated effective regularizers for the low-rank temporal graph reconstruction, which approach was named as **NCTRC** [72].

Moreover, we further improve our NCTR framework and propose the graphical model approach—**NCTRG**. First, we present the low-rank temporal graph reconstruction as a probabilistic graphical model. The graphical model intuitively demonstrates the components (and their statistical relationship) in the graph reconstruction framework. The graphical model also incorporates appropriate distribution priors to avoid over-fitting issues. Second, although the temporal graph used in our preliminary work *NCTRC* naturally embedded the campaign participating frequencies of the business customers, *NCTRC* factorized these frequencies without careful probabilistic justification. To this end, we factorize the campaign participating frequencies with Poisson distribution and integrate the frequency factorization with the low-rank reconstruction of the temporal graphs. In other words, our recommender system based on the temporal graph is backward compatible with the conventional static recommender systems. Finally, the experimental results on real-world B2B marketing data show that the proposed method can capture the unique characteristics of B2B marketing campaign behaviors and effectively improve the quality of the campaign recommendations for challenging B2B marketing tasks.

The remainder of the chapter is organized as follows. In Section 2.2, we formulate the problem of "Next Campaign To Run" (NCTR) with intuitive examples. In Section 3.2, we introduce the main framework of NCTR with Section 2.3.1 defining the temporal graph to represent the buying process, Section 2.3.2 giving an overview on recommendation process, and NCTRC Section 2.3.3 and NCTRG Section 2.3.4 predicting the unobserved graph edges with low-rank graph reconstruction as a probabilistic graphical model. Then, the graph reconstruction is regularized in Section 2.4 with community structure of the business customers. The details of the learning algorithm for the regularized low-rank graph reconstruction are presented in Section .2, after which we present the experimental results in Section 2.6 on several real-world



(a) The campaign sequences.

Code	Campaign Type	Icon
A	Global Search	
B	Email Campaign	
C	Webinar	
D	Trial Download	
E	Trade Show	

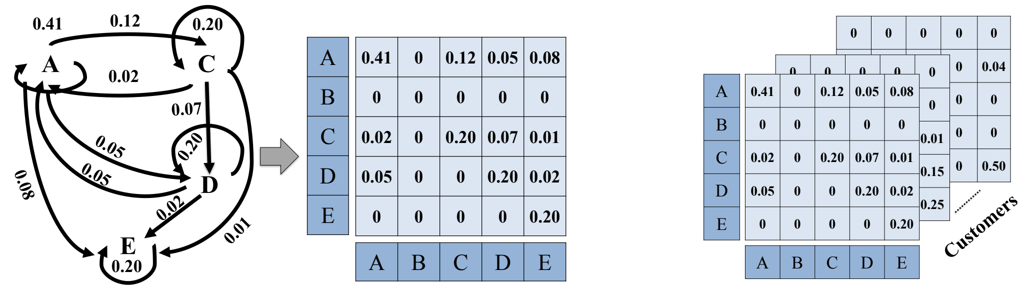
(b) The marketing campaigns.

Figure 2.1: Examples of Customer Behavior Records.

B2B marketing data sets. In addition, we provide an overview of the related works in Section 4.2 and a discussion of the results in the Section 2.8. Finally, the summary of this chapter is in Section 2.9.

2.2 Problem Formulation

We recommend the "Next Campaign To Run" (NCTR) for the business customers by modeling their historical behavior records in the buying processes. Figure 2.1 shows some examples of the behavior records. Specifically, there are three campaign sequences for three customers $C1, C2$, and $C3$ from two companies. For the campaign sequences, we also have the event-happening time when the customer participated in the campaign, therefore we are able to compute the intervals between consecutive events. The intervals can reveal interesting relationship (e.g., the temporal correlations) between the dependent marketing campaigns. For example, both customer $C1$ and $C2$ downloaded the trial product two or three days after attending a webinar. If



(a) Example of transformation of the raw sequence into temporal graph.

(b) Personalized temporal graph.

Figure 2.2: Examples of Personalized Temporal Knowledge Representation.

these similar patterns are followed by the majority of the customers, we could recommend to download trial products to the customers who attended webinars about 2-3 days ago. Intuitively, recommending the dependent campaigns according to the customer's current context in the decision-making process will expedite the buying cycle. Therefore, *our main objective is to exploit the temporal patterns in the behavior records of all customers for providing accurate marketing campaign recommendations.*

2.3 A NCTR Recommender System

This section presents our NCTR recommender system, including: 1) A novel graph-based representation to encode the temporal information in the customer buying process; 2) NCTRC—a low-rank graph reconstruction approach to predict the unobserved graph edges which can be used for NCTR recommendations; 3) NCTRG—the extension of NCTRC with a probabilistic graphical model. 3) A regularization of the graph reconstruction to incorporate the community structure of the customers; 4)

Symbol	**Description**
N, M	The number of customers and campaigns, respectively.
$s^n = (s_1^n, s_2^n, \cdots, s_{L_n}^n)$	The behavior records or buying process of n-th customer.
L_n	The length of campaign sequence.
t_l^n	The campaign-participating time of s_l^n.
G^n	The temporal graph of the n-th customer.
R_{ij}^n	The graph edge from the i-th campaign to the j-th campaign for the n-th customer temporal graph.
R_{jj}^n	The campaign participating frequency of campaign j in the sequence s^n.
$A \in \mathbb{R}^{N \times K}$	The reconstruction coefficients in low-rank temporal graph construction.
$B^k \in \mathbb{R}^{M \times M}$	The adjacency matrix of the k-th graph basis, for $k = 1, 2, \cdots, K$.

Table 2.1: Mathematical Notations.

A stochastic gradient descent learning algorithm to optimize the regularized graph reconstruction.

2.3.1 Temporal Knowledge Representation

Our first step to develop the NCTR recommender system is to design the informative representation of the temporal knowledge hidden in the behavior records of each customer. Inspired by C. Liu et al. [42], we propose the personalized temporal graph which effectively encodes the temporal relationships of the campaigns participated in by each customer.

Suppose we have M campaigns under study. For one specific customer, e.g., the n-th customer, we have his/her behavior records which are represented as a sequence of campaigns $s^n = (s_1^n, s_2^n, \cdots, s_{L_n}^n)$, where $s_l^n \in \{1, 2, \cdots, M\}$ is the l-th campaign in the sequence. We also record the campaign-participating time t_l^n for s_l^n. With these notations, we define the personalized temporal graph G^n for the n-th customer, with all the M campaigns as graph nodes. The direct edge from the i-th node to the j-th node is weighted by:

$$R_{ij}^n = \frac{1}{L_n} \sum_{1 \leq p \leq q \leq L_n} [s_p^n = i \wedge s_q^n = j] \kappa(t_q^n - t_p^n), \tag{2.1}$$

where $\kappa(\cdot)$ is a non-increasing function.

The non-increasing property of the function $\kappa(\cdot)$ enables us to compute a higher edge weight R_{ij}^n if the i-th and j-th campaigns appear close to each other in s^n. For example, we can use the simple Iverson bracket:

$$\kappa(\delta|\Delta) = [\delta \leq \Delta], \tag{2.2}$$

where Δ is a threshold. In this way, we assume the marketing events happened within the temporal range of Δ are temporally related for this customer. An appropriate value of Δ can be thus determined with the domain knowledge in a particular application. More generally, we can also use a smooth function to further discriminate different temporal intervals between the events. In this chapter, we use the truncated exceedance of the Exponential distribution:

$$\kappa(\delta|\Delta, r) = \begin{cases} \exp(-\delta/r) & \delta \leq \Delta \\ \\ 0 & \delta > \Delta \end{cases}. \tag{2.3}$$

Here, we exclude the weight computing between events with relatively large time interval, e.g., larger than Δ, and a scaling parameter r is used to compute the remaining weights. We use this definition for three reasons:

- According to the weight definition in Equation 2.1, the frequency of campaign i in the sequence s^n is included in R^n_{ii} (i.e., i-th diagonal entry), normalized by the sequence length L_n. These frequencies have been used in the design of conventional static recommender systems, while the temporal graphs extend the static frequencies with temporal correlations/dependencies.

- When $r \to +\infty$, Equation 2.3 and 2.2 are equivalent:

  $$\lim_{r\to\infty} \kappa(\delta|\Delta, r) = \kappa(\delta|\Delta).$$

  The reason is that, when r is sufficiently large, each pair of events in the sequence s^n within the temporal range of Δ will be equally connected and weighted in the temporal graph G^n.

- When $r \to 0+$, we have:

$$\lim_{r \to 0+} \kappa(\delta|\Delta, r) = \begin{cases} 1 & \delta = 0, \\ 0 & \delta > 0. \end{cases}$$

  In this case, the graph weight matrix R^n with $\kappa(\cdot)$ defined in Equation 2.3 is almost diagonal, since very few distinct events happened at exactly the same time.

For the sake of simplicity, in the remaining of this chapter, we let $\kappa(\delta) = \kappa(\delta|\Delta, r)$. Table 2.1 lists some of the notations used in this chapter.

To provide an intuitive understanding, 1 shows the computation details of the personalized temporal graph. As can be seen, the graph-based representation translates the event sequences into the pairwise relationships, which captures the temporal closeness between any pair of campaigns. In the following, we utilize the personalized temporal graphs in our NCTR recommender system.

Example 1 *We consider customer $C1$ in Figure 2.1. There are $L_n = 5$ behavior records. We let the scaling parameter $r = (1 + 2 + 3 + 2)/4 = 2$ and Δ is set to* 90 *days, then the personalized temporal graph R^{C1} can be constructed as shown in Figure 2.2a. The following are the calculations for the first row of the adjacency* *matrix:*

$$
\begin{aligned}
R_{AA}^{C1} &= \frac{1}{5}(\exp(0) + \exp(0) + \exp(-\frac{6}{2}) = 0.41, \\
R_{AB}^{C1} &= 0, \\
R_{AC}^{C1} &= \frac{1}{5}\exp(-\frac{1}{2}) = 0.12, \\
R_{AD}^{C1} &= \frac{1}{5}\exp(-\frac{3}{2}) = 0.05, \\
R_{AE}^{C1} &= \frac{1}{5}(\exp(-\frac{8}{2}) + \exp(-\frac{2}{2})) = 0.08.
\end{aligned}
$$

Backward Compatibility. As aforementioned, R_{ii}^n includes the normalized frequency of campaign i in the sequence s^n. These campaign participating frequency can be used to infer the "latent interest" of each business customer as in the conventional recommender systems. Therefore, our approach can be deemed a proper generalization of the conventional static recommender systems considering only the event frequencies as the implicit preferences/rating.

2.3.2 Recommendation with Temporal Graph

Suppose we have constructed the personalized temporal graphs for all customers with sufficient observations. Then for a specific customer with the last campaign i in his/her behavior records s^n, we can sort the campaigns $j = 1, 2, \cdots, M$ and $j \neq i$ with respect to the values $R_{jj}^n \times R_{ij}^n$ in descending order. Here the two terms R_{jj}^n and R_{ij}^n computes the interest preference and the temporal preference, respectively. The campaigns ranked at the top will be recommended to the customer. However, it is expected that the constructed temporal graphs are very sparse with many edges unobserved, and the NCTR recommendation tasks rely on accurate prediction of the unobserved edges. In the next subsection, we develop the collaborative low-rank graph reconstruction approach to predict these unobserved graph edges.

2.3.3 Low-Rank Graph Reconstruction with NCTRC

Inspired by the popular matrix factorization [36, 50] for predicting the unobserved customer-item ratings, we develop the low-rank graph reconstruction approach named as NCTRG for predicting the unobserved edges in the personalized temporal graphs. The assumption is that, each observed temporal graph can be reconstructed by optimally combining a set of graph basis. To be specific, suppose we have N customers and constructed the temporal graph G^n for each $n = 1, 2, \cdots, N$. As introduced in Section 2.3.1, each graph G^n is associated with the adjacency matrix $R^n \in \mathbb{R}^{M\times M}$, where M is the number of campaigns offered by the company. To reconstruct G^n, we assume there are K graph basis and each base graph is associated with a adjacency matrix $B^k \in \mathbb{R}^{M\times M}$ for $k = 1, 2, \ldots, K$. Then we use the graph basis to approximate the adjacency matrix R^n:

$$R^n \sim \sum_k A_{nk} B^k, \tag{2.4}$$

where A_{nk} is the reconstruction coefficients. Note that, the number of graph basis, K, can be deemed the rank of the graph reconstruction, which is set to be much smaller than the number of observed temporal graphs, N: $K \ll N$.

To compute the optimal graph basis B^k for $k = 1, 2, \cdots, K$, and at the same time the reconstruction coefficients in matrix A for all the observed temporal graphs G^n,

$n = 1, 2, \cdots, N$, we can minimize the following reconstruction error:

$$\mathcal{J}(A, B) = \frac{1}{2} \sum_{n=1}^{N} \| R^n - \sum_{k=1}^{K} A_{nk} B^k \|_F^2, \tag{2.5}$$

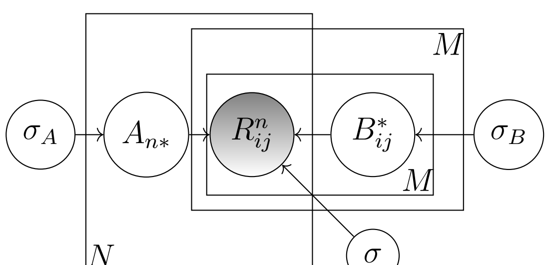
where $\| \cdot \|_F$ denotes the Frobenius norm and we have constraints $A \geq 0$ and $B^k \geq 0$ for all k.

However, due to the sparsity of the temporal graphs, it is more efficient that we compute the reconstruction error with only the observed edges:

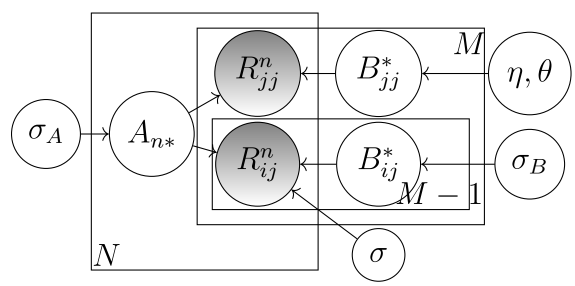
$$\mathcal{J}(A, B) = \frac{1}{2} \sum_{n=1}^{N} \| (R^n - \sum_{k=1}^{K} A_{nk} B^k) \odot I^n \|_F^2, \tag{2.6}$$

where $\odot$ is the Hadamard product operator, i.e., element-wise multiplication of matrices. The binary indicator $I_{ij}^n = 1$ if and only if we have the edge from i to j in graph G^n, i.e., $R_{ij}^n > 0$. Otherwise $I_{ij}^n = 0$.

2.3.4 Low-Rank Graph Reconstruction with NCTRG



(a)



(b)

Figure 2.3: The graphical model of graph reconstruction. The left panel directly adapts PMF for graph reconstruction. The right panel distinguishes the modeling for campaign participating frequency R_{jj}^n and campaign order preference R_{ij}^n.

More intuitively, the low-rank graph reconstruction can be demonstrated as a probabilistic graphical model shown in Figure 2.3. In the left panel, we decompose each graph edge R^n_{ij} by two factor vectors: 1) the reconstruction coefficients in A_{n*}; 2) the edge weights in the graph basis B^*_{ij}, such that:

$$R^n_{ij} \sim Gaussian(\langle A_{n*}, B^*_{ij} \rangle, \sigma),$$

where $\langle A_{n*}, B^*_{ij} \rangle = \sum_k A_{nk} B^k_{ij}$. However, the campaign participating frequencies R^n_{jj} often follows the Poisson distribution instead of Gaussian [13, 39, 40]. Therefore in the right panel of Figure 2.3, we further distinguish the modelling for campaign participating frequency R^n_{jj} and the campaign order preferences R^n_{ij}, $i \neq j$:

$$R^n_{ij} \sim Gaussian(\langle A_{n*}, B^*_{ij} \rangle, \sigma), \ \forall n, i \neq j$$
$$R^n_{jj} \sim Poisson(\langle A_{n*}, B^*_{jj} \rangle), \ \forall n, j$$

The graphical models are flexible to incorporate priors of the latent reconstruction coefficients and the graph basis to reduce the generalization errors (mistakes on unseen data). We use the following priors:

$$A_{nk} \sim Gaussian(0, \sigma_A), \ \forall n, k$$
$$B^k_{ij} \sim Gaussian(0, \sigma_B), \ \forall k, i \neq j$$
$$B^k_{jj} \sim Gamma(\eta, \theta), \ \forall k, j$$

In particular, we use the Gamma distribution $B^k_{jj} \sim Gamma(\eta, \theta)$ since it is the conjugate one with Poisson.

With the above settings, we have the joint probability density:

$$\begin{aligned}
&\Pr(R|A,B)\Pr(A)\Pr(B)\\
=&\prod_{n,i,j\neq i}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(R_{ij}^n-\langle A_{n*},B_{ij}^*\rangle)^2}{2\sigma^2})\right)^{I_{ij}^n}\\
&\times\prod_{n,j}\left(\frac{(\langle A_{n*},B_{jj}^*\rangle)^{R_{jj}^n}}{\Gamma(R_{jj}^n+1)}\exp(-\langle A_{n*},B_{jj}^*\rangle)\right)^{I_{jj}^n}\\
&\times\prod_{n,k}\frac{1}{\sqrt{2\pi}\sigma_A}\exp(-\frac{(A_{nk})^2}{2\sigma_A^2})\\
&\times\prod_{k,i,j\neq i}\frac{1}{\sqrt{2\pi}\sigma_B}\exp(-\frac{(B_{ij}^k)^2}{2\sigma_B^2})\\
&\times\prod_{k,j}\frac{\theta^\eta}{\Gamma(\eta)}(B_{jj}^k)^{\eta-1}\exp(-\theta B_{jj}^k)
\end{aligned}\tag{2.7}$$

where $\Gamma(\cdot)$ is the gamma function ($\Gamma(n+1)=n!$ for non-negative integer n), and I is the indicator such that $I_{ij}^n=1$ if and only if $R_{ij}^n>0$, otherwise $I_{ij}^n=0$, for all n,i,j.

Now, we can formulate the (negative) log-likelihood as the objective function to compute the optimal graph basis B^k for $k=1,2,\cdots,K$, and at the same time the reconstruction coefficients in matrix A for all the temporal graphs G^n, $n=1,2,\cdots,N$:

$$\begin{aligned}
&\mathcal{L}(A,B)=-\log\Pr(R|A,B)\Pr(A)\Pr(B)+\text{const}\\
=&\frac{1}{2\sigma^2}\sum_{n=1}^{N}\sum_{i=1}^{M}\sum_{j\neq i}^{M}I_{ij}^n(R_{ij}^n-\sum_{k=1}^{K}A_{nk}B_{ij}^k)^2\\
&-\sum_{n=1}^{N}\sum_{j=1}^{M}I_{jj}^n(R_{jj}^n\ln\langle A_{n*},B_{jj}^*\rangle-\langle A_{n*},B_{jj}^*\rangle)\\
&+\frac{1}{2\sigma_A^2}\|A\|^2+\frac{1}{2\sigma_B^2}\sum_{k=1}^{K}\sum_{i=1}^{M}\sum_{j\neq i}^{M}(B_{ij}^k)^2\\
&-\sum_{k=1}^{K}\sum_{j=1}^{M}((\eta-1)\log B_{jj}^k-\theta B_{jj}^k)
\end{aligned}\tag{2.8}$$

where $A\in\mathbb{R}^{N\times K}$, $B^k\in\mathbb{R}^{M\times M}$, for $k=1,2,\cdots,K$.

2.4 Community Regularization

Indeed, in addition to modeling the temporal relationships of the marketing campaigns in the complicated decision-making process of the business customers, another important factor which can be leveraged to improve the NCTR recommendations is the community network of the customers. As we mentioned in Section 2.1, in the B2B markets, it is often that multiple customers from the same company will make the business purchase decision together. These customers working on the same buying task or in the same company form a small community, where the members cooperate and communicate with each other. Therefore, during the reconstruction of their temporal graphs, these customers may share similar reconstruction coefficients. To integrate these constraints into our problem formulation, we adopt the so-called community regularization.

Suppose we have the community network encoded in the matrix H, where $H_{uv} = 1$ if and only if the two customers u and v are from the same company, and $H_{uv} = 0$ otherwise. Then our objective function is

$$\mathcal{J}(A, B) = \mathcal{L}(A, B) + \lambda \cdot \Omega(A), \tag{2.9}$$

where the community regularization $\Omega(A)$ encourages the customers from the same company to have similar reconstruction coefficients in A:

$$\begin{aligned}\Omega(A) =& \frac{1}{2}\sum_{u=1}^{N}\sum_{v=1}^{N}\frac{1}{2}H_{uv}\|A_u - A_v\|^2 \\ =& \frac{1}{2}\mathrm{tr}(A'(D-H)A),\end{aligned}$$

where D is the diagonal degree matrix such that $D_{uu} = \sum_{v=1}^{N} H_{uv}$.

The level of the community regularization is controlled by the parameter λ. Specifically, a large λ will tend to make the campaign preferences of different customers to be the same in the same community. On the other hand, a small λ will tend to make the community network effects insignificant. In practice, the optimal λ is dependent on the actual data characteristics, and it can be realized with the cross validation procedure. Moreover, the regularization $\Omega(\cdot)$ is quite flexible to encode different assumptions on the community networks. Generally speaking, the structure in H can also be provided by domain experts or derived from external knowledge on the customer relationships.

2.5 Learning Algorithm

For the sake of simplicity, here we only show the learning algorithm for the NCTRG approach. We use gradient descent procedures to iteratively update the optimization variables: A_{nk}, B_{ij}^k, and B_{jj}^k, for $n = 1, \cdots, N$, $k = 1, \cdots, K$, $i, j = 1, \cdots, M$, and $i \neq j$. The element-wise gradients are given below, but note that the gradient updating can be vectorized for improved efficiency (e.g., in MATLAB/Numpy).

$$
\begin{aligned}
\frac{\partial \mathcal{J}(A,B)}{\partial A_{n_0,k_0}} =& \frac{1}{\sigma^2} \sum_{i=1}^{M} \sum_{j \neq i}^{M} I_{i,j}^{n_0} (R_{i,j}^{n_0} - \sum_{k=1}^{K} A_{n_0,k} B_{i,j}^{k})(-B_{i,j}^{k_0}) \\
& - \sum_{j=1}^{M} I_{j,j}^{n_0} (\frac{R_{j,j}^{n_0} B_{j,j}^{k_0}}{< A_{n_0,*}, B_{j,j}^{*} >} - B_{j,j}^{k_0}) \\
& + \frac{1}{\sigma_A^2} A_{n_0,k_0} + \lambda \frac{\partial \Omega(A)}{\partial A_{n,k}}
\end{aligned}
\tag{2.10}
$$

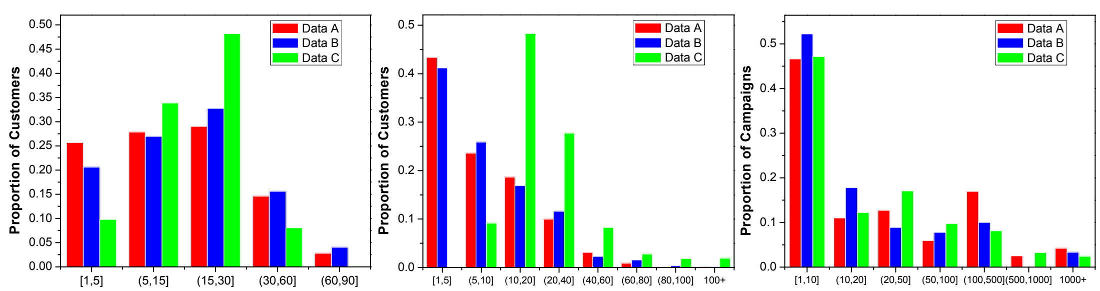
$$\begin{aligned}\frac{\partial \mathcal{J}(A,B)}{\partial B^{k_0}_{i_0,j_0}} =& \frac{1}{\sigma^2}\sum_{n=1}^{N} I^n_{i_0,j_0}(R^n_{i_0,j_0} - \sum_{k=1}^{K} A_{n,k}B^k_{i_0,j_0})(-A_{n,k_0}) \\ &+ \frac{1}{\sigma_B^2}B^{k_0}_{i_0,j_0}\end{aligned} \tag{2.11}$$

$$\begin{aligned}\frac{\partial \mathcal{J}(A,B)}{\partial B^{k_0}_{j_0,j_0}} =& \sum_{n=1}^{N} I_{j_0,j_0}(\frac{R^n_{j_0,j_0}A_{n,k_0}}{< A_{n,*}, B^*_{j_0,j_0} >} - A_{n,k_0}) \\ &+ \frac{\eta-1}{B^{k_0}_{j_0,j_0}} - \theta\end{aligned} \tag{2.12}$$

2.6 Empirical Experiments

In this section, we evaluate the performances of our approach in comparison with several state-of-the-art methods. All the experiments are performed on a GNU/Linux system with 2 CPUs (AMD 2.4GHz) and 4G RAM.

2.6.1 Data Description



(a) Average time interval (b) Sequence length (c) Campaign Popularity

Figure 2.4: Statistics of the B2B Marketing Data.

We have collected three data sets of behavior records of the business customers interested in different products offered by a Fortune 500 company. For each customer, we collect the marketing campaigns participated in by the customer. The campaigns are ordered by the event-happening time as a marketing event sequence. A summary

Characteristics	Data A	Data B	Data C
# of Customers	2,119	568	930
# of Companies	597	250	144
# of Distinct Campaign	116	90	125
Total Campaign Events	24,125	6,758	23,977
Average Time Interval (day)	17.41	19.49	16.88
Average Sequence Length	11.43	11.90	25.78
Average Size of The Communities	3.55	2.27	6.46

Table 2.2: Characteristics of the B2B Marketing Data.

of all the data sets is shown in Table 2.2 with more statistics in Figure 2.4. One can see diverse characteristics in terms of data size (e.g., number of customers, number of companies, and number of campaigns) and event frequency (e.g., total campaign events, average time interval, and average sequence length).

2.6.2 Evaluation Metrics

We use the following evaluation metrics to measure the recommendation performances:

- **Normalized Discounted Cumulative Gain (NDCG)**: The NDCG measures the ranking quality of the recommended list based on a graded relevance scale. It is widely used in researches on recommendation, information retrieval, search engine, etc. For a ranking list with K items:

$$DCG@K = \sum_{k=1}^{K} \frac{2^{rel_k} - 1}{log(k+1)},$$
$$NDCG@K = \frac{DCG@K}{IDCG@K},$$

  where $IDCG$ is the maximum possible DCG for the recommended items, and rel_i is the graded relevance of the list at position i. The range of NDCG is $[0, 1]$, with 1 representing the perfect ranking quality.

- **Precision and Recall**: For a ranking list with K items:

$$Precision@K = \frac{\#relevant\ recommendations}{K},$$
$$Recall@K = \frac{\#relevant\ recommendations}{\#all\ relevant\ items}.$$

  The value of $Precision@K$ and $Recall@K$ closer to 1.0 means better recommendation performances.

- **Area Under the Precision and Recall Curve (AUPR)**: The AUPR usually is computed as the average of all precisions at evenly spaced recall levels [63]. The value is between 0 and 1 and higher is better.

In our experiments, we use the first 65% of the behavior records of each customer for training and the remaining 35% for testing. All the metrics are computed for each customer and then the overall average is summarized to compare different methods or parameter settings. We also report the running time of all methods and the number of iterations for methods using iterative optimization, with the same level of tolerance (a lower bound on the change in the value of the objective function during an iteration) as 0.001.

2.6.3 Benchmark Methods

We compare our approaches (NCTR, NCTRC and NCTRG) with several benchmark methods, including both conventional static recommender systems (Customer/Item Mean, NMF) and state-of-the-art methods based on temporal behavior patterns (FPMC, PIMF) or sequential rule mining (RuleGrowth). We summarize all the methods as follows:

1. **Customer Mean (CustMean)**: This method makes prediction based on the mean value of each customer:
$$\widehat{R}_{ij}^{n} = \frac{\sum_{ij} R_{ij}^{n} I_{ij}^{n}}{\sum_{ij} I_{ij}^{n}}.$$

2. **Item Mean**: This method makes prediction based on the mean value of each item:
$$\widehat{R}_{ij}^{n} = \frac{\sum_{n} R_{ij}^{n} I_{ij}^{n}}{\sum_{n} I_{ij}^{n}}.$$

3. **NMF**: NMF is a widely used collaborative filtering approach [30], which factorizes the customer-campaign binary matrix (1 means the customer participated in the campaign, 0 otherwise).

4. **FPMC**: The Factorized Personalized Markov Chains model [57] transforms the sequential data of each user into a transition matrix and then predict the users

next action by factorizing the matrices.

5. **PIMF**: The Purchase Interval based Matrix Factorization model [77] incorporates purchase interval into the marginal utility model to predict the most relevant items for recommendation at a given time.

6. **RuleGrowth**: The RuleGrowth [24] uses a pattern-growth approach for discovering sequential rules common to several sequences in the sequence databases, which containing sequences of discrete events.

7. **NCTR**: This is our low-rank temporal graph reconstruction approach without any additional regularization.

8. **NCTRC**: This method is our low-rank temporal graph reconstruction approach with the community regularization for customers from the same company.

9. **NCTRG**: This method is the extension of the NCTRC approach with separately defined temporal preferences and campaign interestingness and regularization terms in a graphical model to avoid overfitting.

2.6.4 Performance Comparison and Discussion

In the performance comparison, we are curious about two questions:

- How the enhanced *NCTRG* framework improves baselines (CustMean, ItemMean, NMF) as well as state-of-the-art models (FPMC, PIMF, RuleGrowth)?

- How the enhanced *NCTRG* framework improves our earlier approaches (*NCTR*, *NCTRC*)?

Table 2.3: Performance Comparison on Data A.

Metrics		CustMean	ItemMean	NMF	FPMC	PIMF	RuleGrowth	NCTR	NCTRC	NCTRG
Time(sec.)/#Iteration		**1.949**	8.160	374/199	361/201	6,402/26	21.609	3,952/109	1,963/153	4,013/129
NDCG	@3	0.6353	0.1442	0.7114	0.2971	0.7192	-	0.7538	0.7597	**0.8139**
	@5	0.5890	0.1251	0.6697	0.2711	0.6281	-	0.6937	0.7000	**0.8396**
	@10	0.5333	0.0777	0.6401	0.2603	0.5135	-	0.6549	0.6562	**0.8972**
Precision	@3	0.1884	0.0104	0.2503	0.1408	0.1829	0.1905	0.2543	0.3037	**0.3242**
	@5	0.1339	0.0077	0.1872	0.1135	0.1256	0.2000	0.1858	0.2242	**0.2352**
	@10	0.0926	0.0118	0.1096	0.0744	0.0769	**0.1643**	0.1117	0.1324	0.1450
Recall	@3	0.2933	0.0091	0.3618	0.2871	0.3043	0.0534	0.3672	0.4179	**0.4496**
	@5	0.3308	0.0104	0.4267	0.3878	0.3928	0.0944	0.4331	0.4949	**0.5185**
	@10	0.4261	0.0326	0.4872	0.4992	0.4987	0.1542	0.4973	0.5506	**0.6138**
AUPR		0.0326	0.0001	0.0603	0.0266	0.0417	0.0034	0.0619	0.0843	**0.0941**

Table 2.4: Performance Comparison on Data B.

Metrics		CustMean	ItemMean	NMF	FPMC	PIMF	RuleGrowth	NCTR	NCTRC	NCTRG
Time(sec.)/# Iteration		**0.344**	4.243	38/165	73.5/201	574/16	5.813	445/84	151/70	618/92
NDCG	@3	0.4701	0.3292	0.7114	0.3748	0.7336	-	0.7262	0.7797	**0.8365**
	@5	0.4403	0.1752	0.6697	0.3338	0.6203	-	0.6740	0.8053	**0.8674**
	@10	0.4000	0.1305	0.6302	0.3144	0.5323	-	0.6329	0.8154	**0.9542**
Precision	@3	0.2165	0.0117	0.2832	0.1356	0.1708	0.2667	0.2806	0.3337	**0.3443**
	@5	0.1594	0.0237	0.2000	0.1330	0.1146	0.2333	0.2031	**0.2462**	0.2457
	@10	0.1086	0.0194	0.1211	0.0832	0.0654	**0.2222**	0.1225	0.1423	0.1450
Recall	@3	0.3246	0.0083	0.3916	0.2775	0.3124	0.0636	0.3928	0.4569	**0.5044**
	@5	0.3800	0.0296	0.4561	0.4467	0.4531	0.1303	0.4685	0.5349	**0.5794**
	@10	0.4811	0.0579	0.5314	0.5412	0.5549	0.2242	0.5387	0.5937	**0.6548**
AUPR		0.0420	0.0004	0.0742	0.0255	0.0463	0.0075	0.0740	0.1020	**0.1156**

Table 2.5: Performance Comparison on Data C.

Metrics		CustMean	ItemMean	NMF	FPMC	PIMF	RuleGrowth	NCTR	NCTRC	NCTRG
Time(sec.)/# Iteration		**0.936**	11.075	122/164	214/201	3,108/30	17.625	2,151/131	2,882/140	3,316/109
NDCG	@3	0.1211	0.2887	0.1321	0.4174	0.7332	-	0.7238	0.7380	**0.8206**
	@5	0.1695	0.2504	0.1794	0.3938	0.6621	-	0.6662	0.7015	**0.8521**
	@10	0.1849	0.2271	0.1967	0.3697	0.5661	-	0.6385	0.6721	**0.8942**
Precision	@3	0.0025	0.1446	0.1064	0.2082	0.1787	0.2308	0.3077	0.3304	**0.3753**
	@5	0.0202	0.1243	0.0025	0.1463	0.1226	0.1923	0.2209	0.2537	**0.2852**
	@10	0.0124	0.0943	0.0202	0.0902	0.0673	0.1654	0.1393	0.1463	**0.1818**
Recall	@3	0.0017	0.2158	0.0123	0.3939	0.3783	0.0625	0.4204	0.4547	**0.5001**
	@5	0.0315	0.3029	0.0017	0.4594	0.4658	0.0881	0.4779	0.5423	**0.5952**
	@10	0.0387	0.4307	0.0315	0.5652	0.5673	0.1493	0.5762	0.6131	**0.7094**
AUPR		0.0005	0.0163	0.0022	0.0551	0.0522	0.0061	0.0820	0.1003	**0.1168**

Table 2.3, 2.4, and 2.5 summarizes the comparison results on three B2B marketing data sets, respectively. For all methods, the parameters (if any) are tuned by the cross-validation procedure. More details on parameter selection in our method are provided in Section 2.6.6. From the results, we have the following observations:

1. In general, our methods (*NCTR*, *NCTRC* and *NCTRG*) consistently outperform other baseline methods such as *Customer Mean*, *Item Mean*, and *NMF* on all data sets. This observation well affirms our idea that the performance of NCTR recommendation can be improved by considering the temporal information in customer behaviors and the B2B community structures.

2. *NCTR*, *NCTRC* and *NCTRG* also outperform *FPMC* for predicting the next possible campaigns. The reason is that, the temporal graphs in our methods are more robust in representing the event sequences. For example, in our data, the events happen irregularly with varying time intervals. Meanwhile, there are missing events not recorded in the behavior logs. All of these make the simple transition probabilities of *FPMC* less meaningful. In contrast, our temporal graphs directly compute the temporal correlations between marketing events for recommendations. The irregular time intervals are counted with robust smooth functions and the missing events cannot affect the construction of graph edges between observed events.

3. Moreover, *NCTR*, *NCTRC* and *NCTRG* outperform *PIMF* which considers the diminishing product's utility and user's satisfaction. In our data, the diminishing marginal utility pattern does not fit into the B2B marketing events. For example, a

customer might participate in the same campaign (such as 'Webinar') several times consecutively. Our temporal graphs can model not only this kind of repeatability using self-connecting edges but also temporal correlations between campaigns in the evolving buying processes.

4. In addition, *NCTR*, *NCTRC* and *NCTRG* outperform *RuleGrowh* by a large margin in terms of precision, recall and AUPR. Note that NDCG metric is not applicable for *RuleGrowh* due to the lack of the graded relevance score for the sequential rules. Although *RuleGrowh* as a sequential rule mining approach could find out the most frequently co-occurred campaign sequences, it is less suitable for the B2B marketing scenario than our approaches. This may be due to the reason that *RuleGrowth* is less personalized by not taking into account the individual customer's campaign interests nor the B2B community structure among the customers.

5. *NCTRC* and *NCTRG* achieve higher recommendation quality than *NCTR*, which shows the benefits gained by incorporating the community information when computing the low-rank graph reconstruction. Moreover, our proposed *NCTRG* further improves *NCTRC* significantly. From the results in the three tables, we can observe an average of 0.0660 improvement in terms of all the measures for *NCTRG* over *NCTRC*. This improvement can be attributed to the following reasons. First, the Poisson-based enhanced method is more appropriate for modeling the campaign frequency to capture the campaign interestingness preference. Second, the *NCTRG* separately models the campaign frequency and the temporal preference in the temporal graph, in which providing a more rigorous model than the *NCTR*

and *NCTRC* applied an approximation solution. Moreover, the added regularization terms in the model also help to avoid overfitting thus to improve the overall performance.

Comparisons across different datasets. In general, we find that consistent improvements of the proposed extension over all the baseline methods, even though these three datasets differ in terms of data size, campaign sequence length, and community density. We also notice some interesting different results of these three datasets. First, we find that in Data A and B, the method CustomerMean outperforms ItemMean, while in Data C, the method ItemMean outperforms CustomerMean. This can be explained by the statistics in Figure 2.4 and Table 2.2. In Data C, the average sequence length is much longer than those in Data A and Data B. In this case, campaign-based approach can generally generate better results since more campaign information has been utilized. In contrast, Data A and Data B have relatively shorter sequence length, this is the possible reason that the ItemMean performed worse than CustomerMean. Secondly, the average size of the B2B community varies in these datasets. Specifically, dataset B has the smaller size of average 2-3 contacts for each company, which dataset C has the larger size with above 6 contacts for each company. Due to the diverse size of community, the improvement of the methods with community regularization also varies. From the results we find that, the smaller size of the community, the better improvements of the approach. The possible reason for the finding is that the smaller community means tighter connection among the community members, thus more significant influences on each others' behavior.

Summary. The temporal preference, community relationships and the campaign interesting frequency are three of the most important characteristics for B2B marketing data, and play an important role in the B2B campaign recommendation. The *NCTRC* which fuses the first two factors–the temporal preference and the community relationships into the B2B customer overall preferences can improve the *NCTR* and other baselines. However, an integrated and enhanced of all three characteristics for the B2B campaign recommendation lead to further improvements. The proposed *NCTRG* not only considers the temporal preferences from the behavior sequence and the community relationship for recommendation, but also constructs the temporal graph by considering the campaign visiting frequency as the campaign interestingness. As a result, we can observe significant improvements over all the baseline algorithms. Also, as shown in the performance comparison between *NCTRC* and *NCTRG*, we observe improvements by *NCTRG*, as Poisson distribution is more suitable for modeling the campaign frequency data.

2.6.5 Case Study

To provide a better understanding, we demonstrate some detailed recommendation results in Table 2.6 and Figure 2.5 for 5 customers from two companies. With different last behavior records, our approach recommends to each customer a list including distinct contents. For example, as shown in Figure 2.5, the last record of customer $C1$ is "Search", which indicates that $C1$ is still in the primitive decision-making stage. Therefore, the recommended campaigns for $C1$ include first "Email Campaign" and "Web Advertising" to enhance the product awareness, and later "Webinar",

Customer	Last Record	Recommended List
		Company A
C1	Search	Email, Web Advertising, Webinar, Seminar, Trial Download
C2	Webinar	Webinar, Seminar, Trade Show, Corporate Event, Conference
		Company B
C3	Trade Show	Training, Trade Show, Webinar, Corporate Event, Outbound Telemarketing
C4	Trial Download	Trade Show, Training, Webinar, Seminar, Corporate Event
C5	Search	Webinar, Seminar, Trade Show, Trial Download, Training

Table 2.6: Examples of Recommended Lists.

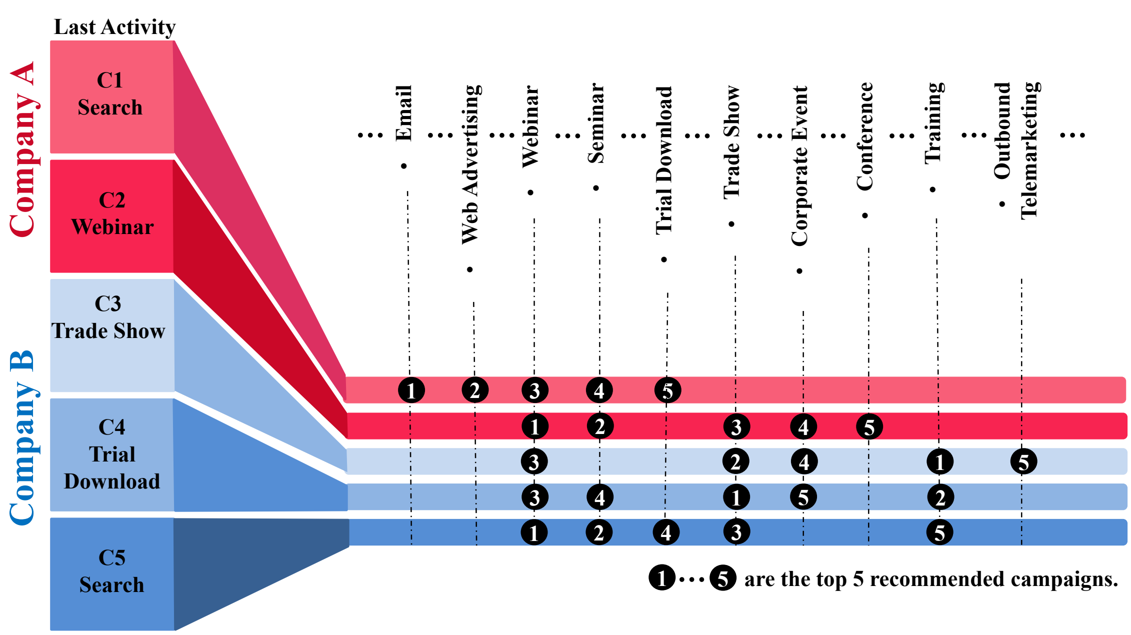


Figure 2.5: Examples of Recommended Lists.

"Seminar", and "Trial Download" to boost the interest level of the customer $C1$. In comparison, with "Trade Show" as the last record, $C3$ is currently in a more mature status toward purchase. Thus, this customer is provided with some late-stage marketing campaigns such as "Training", "Corporate Event", and "Outbound Telemarketing", which can accelerate the buying process for final purchase decision.

Another interesting observation is that, although customer $C5$ has the same last record with $C1$, the customer $C5$ is provided with very different recommendations. This becomes natural with the consideration of the community relationships. Specifically, we investigated the historical records of other community members and found that the customers from this company are more proactive. They are already well aware of the products without those advertisement related campaigns. Therefore, some late-stage campaigns such as "Trade Show" and "Training" are recommended to $C5$ for further information.

Furthermore, by separately modeling the temporal preferences and the campaign interestingness and especially using Poisson distribution to model the campaign participating frequency, NCTRG can better capture customers' preferences and recommend the campaigns with consideration of both time relevance and interestingness. For example, "Webinar" is recommended again to $C2$ even though it is same as the last record. This is due to the fact that "Webinar" is one of the most popular B2B marketing campaign types and $C2$ attended webinars several times.

These observations clearly show that, for recommending the next campaign to run, our approach can model not only the customer-specific behavior preferences but also the community-related behavior commonalities. Moreover, both campaign interestingness and temporal relevance are carefully taken into consideration for recommendation. All these perspectives are effective to improve the recommendation performances in terms of various measures.

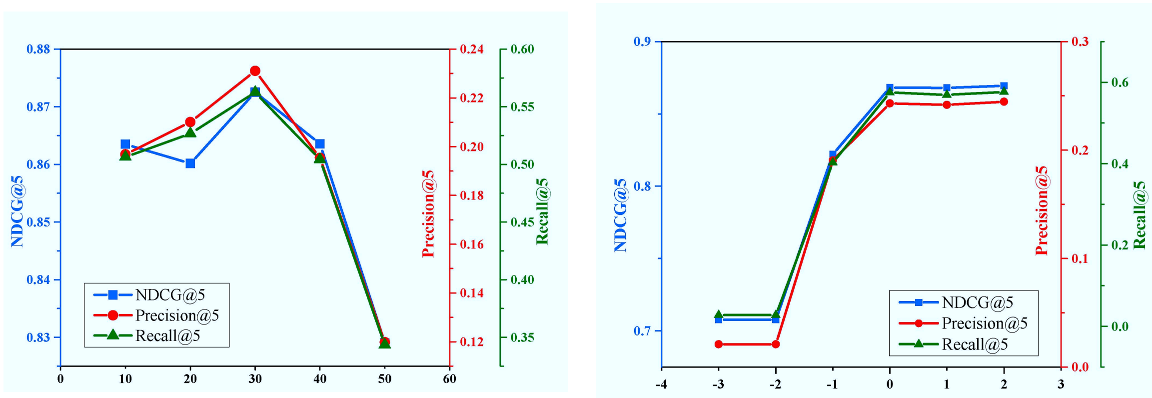
2.6.6 Parameter Selection

Now we consider the important parameters $(\Delta, r, K, \lambda, \sigma_A, \sigma_B, \theta, \eta)$ in our NCTR recommender system and provide detailed discussion on the impacts of parameter changing values on the model performance, which can also provide some guidance for the parameter value selection. Specifically, we optimize the parameters following a nature order, which first decides temporal graph parameters with domain knowledge, then the number of graph bases, the degree of regularization, finally probability prior parameters. This procedure is computationally efficient and produces superior results on our data sets. In general case, an iterative optimization of parameters or an exhaustive grid search should be used for the best performance in practice. For the sake of simplicity, we only show the results for the validation dataset of Data B.

Temporal Graph Parameters: The values of two parameters (Δ and r) for constructing the temporal graphs can be chosen according to the domain knowledge. For example, the thresholding parameter Δ in Equation 2.3 is set to be 90 days in our data sets. The reason is that the customer will make decisions hardly based on actions taken three months ago. Moreover, to make the numerical computing stable in the function $\exp(\cdot)$, the scaling parameter r for computing the temporal correlation in Equation 2.3 can be chosen as the average time interval between all the consecutive marketing events.

The Number of Graph Basis K: We show how the value of graph basis K for reconstructing the temporal graphs impact on model performances in Equation 3.1. As shown in Figure 2.6a, we plot the recommendation performances with increasing number of graph basis. As can be seen, the performances in terms of different measures vary significantly with the different numbers of graph bases. It is worthy to note that the performances might not increase with more bases. The reason is that, more bases imply higher modeling complexity and may lead to overfitting in the training data and decreasing generality of the identified graph bases. Based on Figure 2.6a, we see that $K = 30$ is a feasible trade off between the modeling complexity and the empirical accuracy, with which we achieve the optimal performance consistently in terms of all measures. Therefore we choose $K = 30$ for this data set.

The Community Regularization Parameter λ: Now we discuss the set of regularization parameter λ, which controls the degree of the regularization using the community network in Equation 3.1. The appropriate values of these parameters can also be chosen according to the recommendation performance. Intuitively, if we use a small value of λ, then we only employ the temporal graphs encoding the customer behavior preferences for making recommendations. On the other hand, if λ is larger, the community network information will have a stronger impact on the reconstruction of the temporal graphs. To choose the optimal value for λ, Figure 2.6b shows the recommendation performances with different increasing values of λ. When λ is greater than 0, it becomes stable and gives the optimal results in terms of Precision, Recall and NDCG.



(a) K

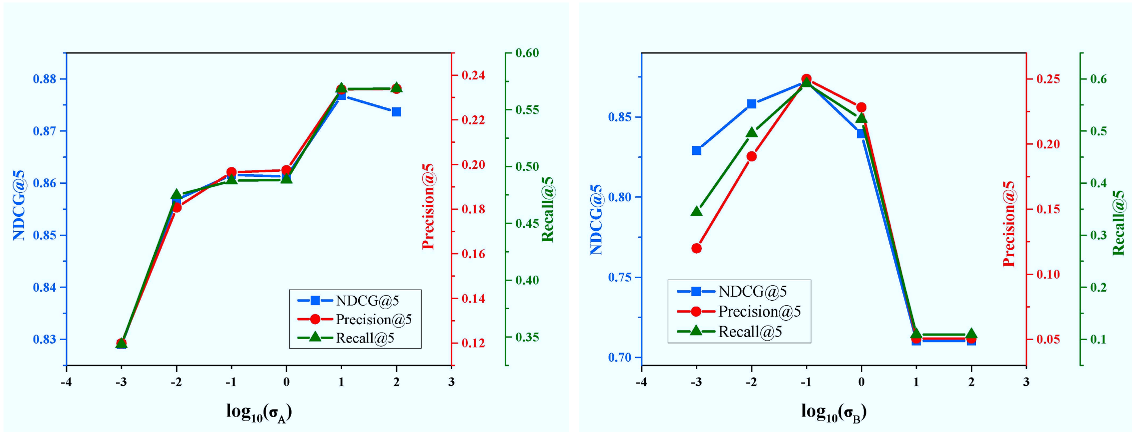
(b) λ

Figure 2.6: Impacts of K and λ on Model Performances.

The Graphical Model Priors Parameters $\sigma_A, \sigma_B, \theta, \eta$: Finally, we demonstrate the impacts of the changing values for the set of graphical model prior parameters σ_A, σ_B, θ and η for different distributions, respectively. As shown in Figure 2.7 and Figure 2.8, the suitable values of these parameters can also be decided based on the recommendation performance. For σ_A, σ_B, and θ, we can find the peak values of the recommendation performances from Figure 2.7a, Figure 2.7b and Figure 2.8a, respectively. Specifically, for Data B, the optimal results of the performance measures can be reached with the parameter values around $\sigma_A = 10$, $\sigma_B = 0.1$, and $\theta = 1$. While for η, the range of η is good for $4 \leq \eta \leq 16$ based on the Figure 2.8b.

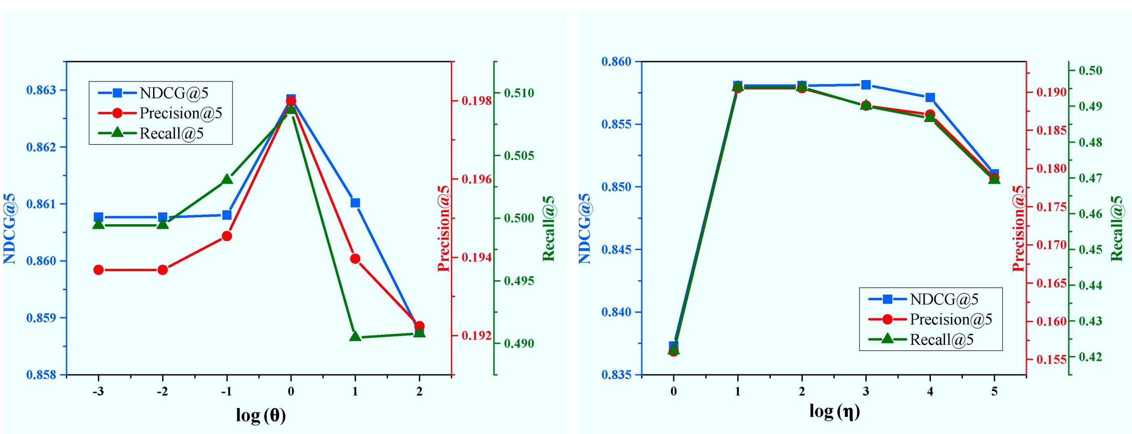
2.7 Related Work

In this section, we reviews several categories of the existing work that are closely related to the proposed research.



(a) σ_A (b) σ_B

Figure 2.7: Impacts of σ_A and σ_B on Model Performances.



(a) θ (b) η

Figure 2.8: Impacts of θ and η on Model Performances.

2.7.1 Traditional Recommender Systems

The research of recommender systems have been an active topic in recent years. In general, there are three major recommendation approaches including content-based methods, collaborative filtering (CF), and hybrid methods [9, 1]. One of the most widely used approaches is the collaborative filtering, which provides recommendations by predicting what users will like based on their similarity to other users. One way to define the similarity is to compute statistics in the user-item rating data. Another way is to define the similarity implicitly by fitting the observed data with latent models which can be used to make unobserved predictions. The model-based approaches received a great attention especially in the Netflix movie recommendation competition [36], where the low-rank matrix factorization was shown effective and efficient with sparse observations [55, 50, 71].

2.7.2 Temporal Recommender Systems

Another type of useful information in the design of recommender systems is the temporal pattern, which is still under-explored yet. For example, [35, 22, 71, 15] showed that the temporal dynamics can be considered in the collaborative filtering model to learn the dynamic characteristics of the users and items. Also, Tang & Zhou [62] proposed to extract dynamic features using time-series analysis and apply the adaptively weighting algorithm to make recommendations. N. N. Liu et al. [43] proposed to combine explicit and implicit user feedbacks to learn the seasonality or short-term preferences for movie recommendations. Xiang et al. [70] proposed to combine the user similarity as the long-term preference and the product similarity as the short-term bias to make recommendations.

Another research direction of the temporal recommender system is to exploit the time factors for short-term "next-item" recommendations. To this end, Rendle et al. [57] integrated the latent factor model and Markov chain model for next-basket recommendation. Wang & Zhang [67] proposed an opportunity model to estimate the follow-up purchase probability of a user at a specific time. Yap et al. [74] proposed to learn user-specific "sequence important knowledge" through personalized sequential pattern mining. Recently, to improve the next-product recommendations in e-commerce, the time interval between purchase behaviors has been modeled by Zhao et al. [77, 78]. Similar ideas have also been used in other applications [12], such as the recommendation of the next-POI (Point-of-Interest) to check-in in location-based services [17, 76]. We implemented the closely related and applicable methods in these researches and discussed more details in Section 2.6.

2.7.3 Recommender Systems with Social Information

Finally, the social information has been used to improve the recommendation performances. In particular, [46, 47, 48] proposed to integrate the social network with the matrix factorization method to learn the latent factors for both users and items with different applications. To improve review quality prediction, Lu et al. [45] developed a generic framework for incorporating social context information by adding regularization constraints in the text-based predictor. To better utilize user's social trust information, X. Yang et al. [73] developed a category-specific social trust circle based model with the user-item rating data combined with social network data. Moreover, in [19], the social correlation is used with Latent Dirichlet Allocation (LDA) to model the users' adoption of items. The authors further devised a hybrid model that combines a user's own latent factors with her friends' for adoption prediction. In [60], a joint personal and social latent factor (PSLF) model is used for social recommendation by explicitly expressing the varieties of the social relationships for each user. In this chapter, we have a simple community structure of the business customers. We observed that the individuals in the same community collaborate with each other and thus we incorporate a community regularization to reflect the relationship in our NCTR approach to improve the marketing campaign recommendations.

2.8 Discussion

Here, we discuss the advantages and limitations of this study. From the experimental results, we can see that the proposed NCTR framework works very well for predicting the B2B customers campaign preferences by exploiting the temporal and community characteristics of the B2B marketing campaign data. Also, in this paper, we describe the work in a domain depended (i.e., B2B marketing) way where users are B2B contacts from different companies, items are various marketing campaigns, and features of items are extracted from the campaign behavior sequences and so on. However, it is worth noting that the idea of integration the temporal factors and the relationships of the users into a recommendation framework should be generally applicable to other recommendation scenarios.

In the meantime, the NCTR framework has some limitations. First, we focus on designing the recommendation algorithm for the next possible campaign based on the customer's past behavior records. So there are maybe some cold-start problems for a new customer come in and without any behavior records. Thus, our approach may be more useful are existing customers with a few behavior records. Also, if we want to deploy this work for real-world services, we have to incorporate more practical functions. Second, there are some limitations with the performance evaluation, which is justified based on the three test data and a simple user study. For instance, preferred and relevant campaigns in the test set may be just a small fraction of the entire relevant ones that are actually favored by the customers. For real-world applications, more sophisticated experiments are required. In addition, note that, currently we define the communities for customers from the same company. In the future, we may identify more fine-grained community structures among customers, e.g., customers can form different groups for different buying tasks.

2.9 Summary

In this chapter, we developed a novel recommender system to combine the temporal and social factors captured in customer behavior records and the customer community networks for B2B marketing campaign recommendations. The goal is to provide the marketers with a better marketing strategy to expedite the customer conversion cycle and boost the customer conversion ratio. The proposed B2B marketing campaign recommender system strategically integrates the temporal preferences, community relationships and the campaign interestingness in the framework. Specifically, we first represented the rich temporal content in customer behavior records using the temporal graphs. Next, with the personalized temporal graphs, we computed the low-rank graph reconstruction to predict the unobserved graph edges. Moreover, we regularized the graph reconstruction with the community network of the business customers. Furthermore, the proposed approach *NCTRG* extended the preliminary work *NCTRC* by considering the skewed campaign frequency data characteristic of B2B marketing data, to incorporate the campaign interestingness information in the model. Finally, we developed efficient algorithms to compute the optimal solutions, which we have applied on several real-world B2B marketing data sets. The experiments clearly validated the effectiveness of the proposed approach in comparison with the state-of-the-art methods.

CHAPTER 3

BUYER TARGETING OPTIMIZATION: A UNIFIED CUSTOMER SEGMENTATION PERSPECTIVE

In marketing analytics, customer segmentation (clustering) divides a customer base into groups of similar individuals, while buyer targeting (classification) identifies promising customers. Both customer segmentation and buyer targeting help the business to improve marketing performances by allocating resources to the most profitable customers. Due to the heterogeneity across the customer groups, some studies have been made on combining the tasks of customer segmentation and buyer targeting for tailored marketing strategies. However, these efforts usually combine these two tasks in a simple step-by-step approach. It is still unclear how to implement these two tasks in a more integrated and optimized way, which is the research objective of this work. Specifically, we formulate customer segmentation and buyer targeting as a unified optimization problem. Then, the customer segments are adaptively realized during the targeting optimization process. In this way, the integrated approach not only improves the buyer targeting performances but also provides a new perspective of segmentation based on the buying decision preferences of the customers. The unified customer segmentation and buyer targeting method not only quantifies the purchase tendency of a specific customer but also characterizes the buying decision behaviors at the segment level. We also develop an efficient *K-Classifiers Segmentation* algorithm to solve the unified optimization problem. Moreover, we show that the customer segmentation based on the buying decision preferences can also be consistent with the features on customer profiles. Finally, we have performed the extensive experiments on several real-world Business to Business (B2B) marketing data sets. The results show that our approach offers not only more accurate targeting of promising customers but also meaningful customer segmentation solutions with interpretable buying decision preferences for each customer segment.

3.1 Introduction

Customer segmentation and *buyer targeting* are two intelligent components of the customer relationship management (CRM) systems. Specifically, customer segmentation targets on dividing the customer base into groups of individuals who share similar profiles, product needs, or marketing priorities. Customer segmentation provides better understanding of the customers' characteristics at a finer granularity level and enables differentiated marketing strategies to meet the customers' needs. On the other hand, buyer targeting identifies promising customers and allocates marketing resources on them to increase sales/profits. Finding better ways of customer segmentation and buyer targeting is essential to reduce marketing cost and boost business performances in modern marketing analytics [2, 3, 18].

Marketing professionals traditionally accomplish these two tasks in two independent steps: using a clustering method for customer segmentation and using a classification method for buyer targeting. Given the discrepancies among customer groups, the segmentation results have been leveraged to improve the classification performance for buyer targeting [51, 18, 3, 58]. For example, Chou et al. [18] proposed to first use K-Means clustering to segment customers and then build the segment-wise predictive models for better targeting the promising customers. Also, Apte et al. [3] developed individually tailored predictive models for each segment to maximize targeting accuracy in the direct-mail industry. In such a step-by-step approach, the buyer targeting (the second step) becomes dependent on the results of customer segmentation (the first step). However, the customer segmentation has to be implemented independently and can only provide limited improvements for the subsequent buyer targeting.

Is it possible to further optimize the buyer targeting performances by implementing these two tasks in a more integrated way? This is the research question we would like to answer in this chapter. To this end, we investigate how to integrate these techniques in a unified optimization framework. Our key idea is to group the customers based on their decision preferences, which are quantified with the targeting models. In this way, it is possible to divide the customer base in an optimal way with respect to the targeting performance. Also, segmentation and targeting in our unified approach are intrinsically related and can be mutually supportive. To the best of our knowledge, the integration of both segmentation and targeting into a unified and optimized process is an innovative approach in solving classic marketing problems. In a more general sense, our approach unifies the clustering process and the classification modeling so that the clustering solution can best boost the classification performance. Our algorithm can also be used for other applications in addition to marketing optimizations.

More specifically, we first provide a novel mathematical formulation to integrate the customer segmentation and the buyer targeting into a unified optimization problem. Inspired by the K-Means clustering, we then develop an iterative algorithm (*K-Classifiers Segmentation*) to optimize the customer segmentation and targeting simultaneously. In comparison with K-Means clustering using the centroid as the partition and update criteria, we learn a set of targeting models and assign each customer to his/her most appropriate model. The customers assigned to the same targeting model form one customer segment, where the customers' buying decision preferences can be explained by the associated targeting model. As a result, targeting accuracy is improved due to the buying decision oriented segmentation.

To improve the interpretability and the robustness of the results, we further develop a *profile-consistent K-Classifiers Segmentation* algorithm. Indeed, using the straightforward process similar with K-Means clustering, the resultant segmentation may group customers with similar profiles into very distinct segments, which are difficult for marketing professionals to interpret. To solve the profile inconsistency issue, we exploit the Nearest Neighbor Clustering framework [10]. With the profile-consistent algorithm, the identified segmentation is consistent with not only the customer profiles but also the customer decision preferences.

Finally, we demonstrate our approach on both synthetic data set and real-world B2B marketing data sets. We use the synthetic data to better illustrate the algorithm details. The results of the real-world data sets show that the proposed approach can greatly improve the accuracy of the buyer targeting than other benchmark methods. Moreover, we validate that our approach can also offer good clustering performance comparing to K-Means clustering. In addition, the interpretation of the buying decision oriented segmentation result can help us to understand the different behaviors of the customers. The marketers can use the decision preferences of each customer segment to develop tailored marketing campaigns to attract prospects.

In summary, the contributions of this work include:

- An innovative formal framework to integrate the customer segmentation and buyer targeting into a unified optimization problem;
- A practical profile-consistent K-Classifiers Segmentation algorithm to optimize the unified problem with interpretable solutions;
- An empirical study on real-world data, showing promising results on both better targeting accuracy and segmentation with actionable marketing implications.

Overview. The remainder of this chapter is organized as follows. In Section 3.2 we provide a detailed description of our integrated framework, and in Section 3.2.3 we propose a Profile-Consistent algorithm to enhance our approach. Next, Section 3.3 reports the experimental results for both synthetic data set and real-world B2B data set. Section 3.4 shows the related works and finally Section 2.9 concludes this work.

3.2 Buyer Targeting with Unified Segmentation

In this section, we develop the unified optimization framework to address the following two complementary tasks:

1) **Customer Segmentation**: The task of customer segmentation is to find a set of segments, where similar customers are grouped together. A clustering algorithm is often used for this purpose by treating each cluster as one segment.

2) **Buyer Targeting**: With a large number of customers, buyer targeting uses classification algorithms to identify the promising prospects for marketers to pursue.

In the literature, to cope with the heterogeneity of the customers, the buyer targeting models are often learned for each segment separately, and conventionally, the customer segmentation is an independent pre-step. In this work, we show that these two tasks are indeed intrinsically related and can be mutually supportive. We propose a unified framework to simultaneously segment the customers and fit the buyer targeting models. The identified segmentation is consistent with not only the customer profiles but also the customer decision preferences, which are quantified with the loss function of a targeting model.

Specifically, we unify the two tasks as an integrated optimization problem. Suppose we have a data matrix $X \in \mathbb{R}^{N \times D}$, where the n-th row x_n represents the profile features of the n-th customer. Also, we have the responses $\{y_n | n = 1, \cdots, N\}$, with $y_n = +1$ for buyers and $y_n = -1$ otherwise. We want to group the customers into K segments (clusters), $\{S_1, S_2, \cdots, S_K\}$, and learn the buyer targeting model (classifier) C_k for each segment S_k respectively. In the following, we take the linear model as the example, i.e., the decision function of C_k is of the form:

$$C_k(x) = \langle x, h_k \rangle + c_k,$$

where h_k represents the model coefficients and c_k is a constant bias.

With the linear buyer targeting models, we simultaneously group the customers and optimize the model parameters by minimizing the following *total loss*:

$$\begin{aligned}\mathcal{J}(S,C) &= \frac{1}{N}\sum_{n=1}^{N}\text{loss}(x_n, y_n|C_{\ell_n}) \\ &= \frac{1}{N}\sum_{k=1}^{K}\sum_{n\in S_k}\text{loss}(x_n, y_n|C_k)\end{aligned} \tag{3.1}$$

where ℓ_n is the segment (cluster) assignment of the n-th customer, i.e., $\ell_n = k$ if and only if $n \in S_k$. In other words, each customer-specific loss in the total loss is computed with the user's respective targeting model.

This formulation is flexible enough to incorporate different types of loss functions in different classification models. In this work, we consider both Logistic Regression and Support Vector Machine (SVM), while other models can also be applied. To be specific, when applying Logistic Regression, we have the *logit* loss:

$$\text{loss}(x, y|C) = \log(1 + \exp(-y \cdot C(x))), \tag{3.2}$$

and when applying SVM, we have the *hinge* loss:

$$\text{loss}(x, y|C) = \max\{0, 1 - y \cdot C(x)\}. \tag{3.3}$$

Here, x is a customer profile, y is the corresponding response, and C is the buyer targeting model.

3.2.1 K-Classifiers Segmentation

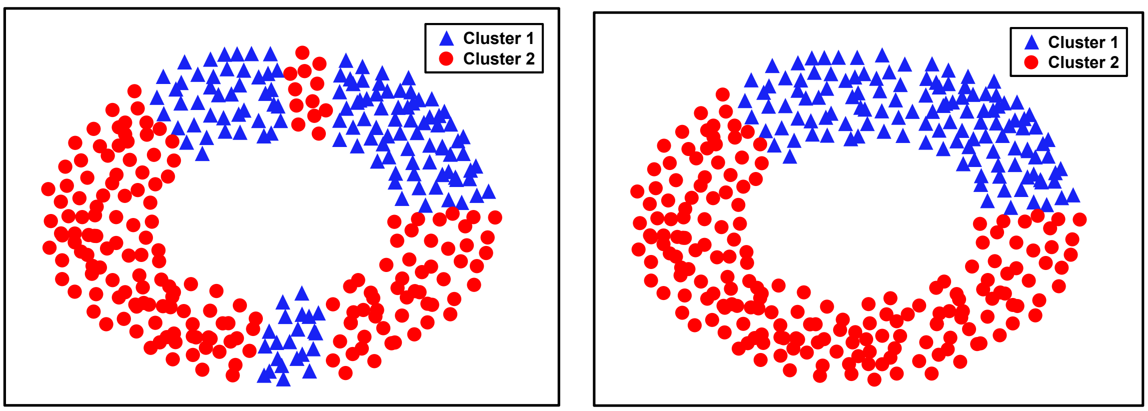
Intuitively, the objective in Equation 3.1 is very similar to that of the K-Means clustering. We replace each centroid in K-Means clustering with a classifier, and we replace the distance between the centroid and a nearby point with the loss of that point in our classification model. Consequently, the problem to minimize $\mathcal{J}(S,C)$

is NP-hard. Starting with a random initialization, we use an iterative process to optimize the segmentation and targeting models (Algorithm 1). The K-Classifiers Segmentation Algorithm begins with the randomly formed (or predefined) segments. Then, the following two steps are iterated until convergence. First, in the update step, we learn a classifier C_k for each cluster S_k. Second, in the assignment step, we assign every point x_n (with response y_n) to the ℓ_n-th classifier with the minimum loss based on the specified loss function.

Although the optimization process is similar with K-Means clustering, the K-Classifiers Segmentation Algorithm does not inherit critical shortcomings of K-Means clustering. Particularly, the simple K-Means is based on Euclidean distances, and thus it is prone to generate spherical clusters with similar sizes. However, the real data sets may not satisfy these assumptions. In contrast, the K-Classifiers Segmentation Algorithm is based on a loss function, which quantifies the customer decision preferences without any assumptions on the shapes or sizes of the segments.

3.2.2 Profile Inconsistency Problem

The optimization in Algorithm 1 works solely with the classification loss of the data. Therefore, in some cases, it may lead to profile inconsistency. In other words, there may be the circumstance that points close to each other may end up being assigned into different clusters. For instance, Figure 3.1(a) shows the situation that both of two clusters (represented by blue triangle and red circle respectively) have points in the associated cluster region but belong to the other cluster (e.g. the red circles in the upper right corner and the blue triangles in the lower right corner), while Figure 3.1(b) shows the desired profile-consistent results. For the case of customer segmentation, the profile inconsistency means that customers with very similar feature profiles may be grouped into different segments, and consequently, it is very difficult to interpret and apply the results in practice. To cope with this challenge, we further propose profile-consistent strategy in the following section.



(a) An Example of Profile Inconsistency Case.

(b) An Example of Profile-Consistent Case.

Figure 3.1: Inconsistency vs. Consistency.

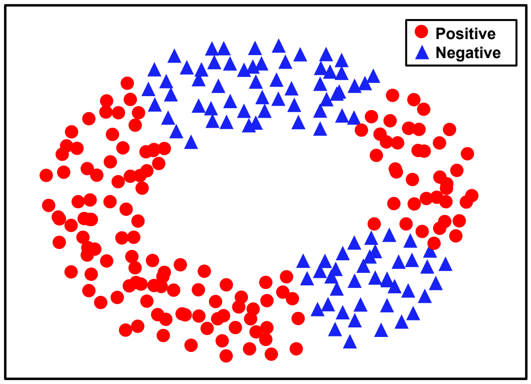
3.2.3 Profile-Consistent Algorithm

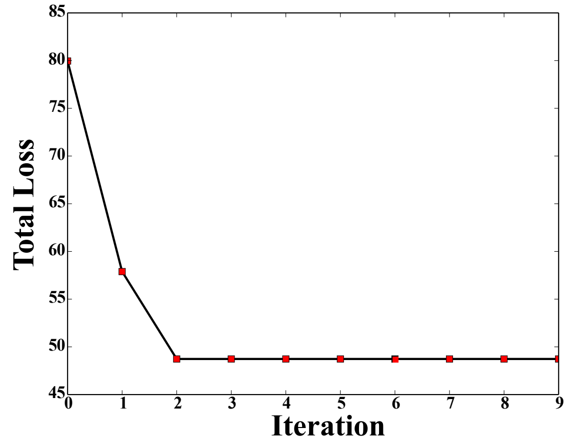
To improve the interpretability and the robustness of the results, now we optimize the $\mathcal{J}(S, C)$ with S_k consistent with the customer profiles $x_n \in S_k$. We adopt the *Nearest Neighbor Clustering* algorithm [10], which is very flexible and shown to produce consistent clustering solutions with arbitrary objective functions. Using the idea of Nearest Neighbor Clustering, the Profile-Consistent K-Classifiers Segmentation is provided in Algorithm 2.

	K-Means	K-Classifiers	Profile Consistent K-Classifiers
Initialization	Randomly select K observations as the initial centroids.	Randomly select K observations as the initial centroids.	Randomly select M seed points to construct M sub-regions.
Update Criteria	The centroid.	The classifier learned based on all the points in the segment.	The classifier learned based on all the points in the segment.
Objective	Minimize the within-cluster distance.	Minimize the point-wise classification loss.	Minimize the classification loss for each sub-region.

Table 3.1: The Comparisons of Three Algorithms.

In addition to settings in Algorithm 1, a new parameter M $(K \leq M \ll N)$ is needed to form sub-regions in the space, and the optimization process is constrained with consistent clustering assignments for each sub-region. The sub-regions are formed with the closeness between the customers, and thus this procedure can improve the profile consistency. To be specific, we randomly select M seed points and construct the Voronoi decomposition as the sub-regions $T_1, \cdots, T_M$. Then, we start with random segment assignments for the sub-regions and iterate the update step and the assignment step as in Algorithm 1. The only difference is that, the assignment step will group the sub-regions instead of individual points. Intuitively, we actually first identify profile-consistent sub-segments of customers, and then identify the final customer segmentation by re-allocating the sub-segments based on the targeting models. In this way, we can simultaneously learn the targeting models and identify the profile-consistent customer segments.





(a) Scatter plot of the synthetic data with two classes (red and blue.

(b) The total loss at each iteration.

Figure 3.2: Synthetic data plot and the loss curve.

To further illustrate the profile-consistent algorithm, in Figure 3.3, we show the intermediate optimization results on a synthetic data set. Specifically, we have generated a two dimensional data set with two classes (the positive class in red and the negative class in blue), as shown in Figure 3.2(a). As can be seen, these two classes (red and blue) cannot be classified using just one simple linear model. However, if we partition the space into several segments, it is possible that the classes are separable within each segment. For example, the data can be divided two segments, one is the top blue points with large group of red points on left, and the other is lower blue points with small group of red points. As illustrated in Figure 3.2(b) and Figure 3.3, the profile-consistent algorithm can successfully identify the two segments and fit their respective classification models by minimizing the total classification loss, in a small number of iteration steps.

In Table 3.1, we summarize the differences among the K-Means clustering (KM), the K-Classifiers Segmentation (KC), and the Profile Consistent K-Classifiers Segmentation (PC) with respective to three specific aspects: the initialization step, the update criteria, and the optimization objective. Again, as we mentioned in Section 3.2.1, due to the intrinsic differences among the KM, our proposed KC and PC algorithms, the KC and PC algorithms will not inherit critical shortcomings of K-Means clustering.

3.2.4 Convergence Analysis

Similar to classical K-Means clustering algorithm, our approaches also have the property of convergence. The total cost monotonically decreases since each iteration of Algorithm 1 and Algorithm 2 necessarily lowers the cost, as shown in the following.

Let $C_1^{(t)}, \ldots, C_k^{(t)}, S_1^{(t)}, \ldots, S_k^{(t)}$ denote the classifiers and clusters at the start of the t-th iteration of both algorithms. The t-th iteration assigns each data point to the classifier with the minimum loss based on specified loss function. Therefore $\mathrm{loss}(S_1^{(t+1)}, \ldots, S_k^{(t+1)}; C_1^{(t)}, \ldots, C_k^{(t)}) \leq \mathrm{loss}(S_1^{(t)}, \ldots, S_k^{(t)}; C_1^{(t)}, \ldots, C_k^{(t)})$. Next, each cluster is reformed by the data points with corresponding classifier, then

$\mathrm{loss}(S_1^{(t+1)}, \ldots, S_k^{(t+1)}; C_1^{(t+1)}, \ldots, C_k^{(t+1)}) \leq \mathrm{loss}(S_1^{(t+1)}, \ldots, S_k^{(t+1)}; C_1^{(t)}, \ldots, C_k^{(t)})$.

3.3 Experimental Results

In this section, we demonstrate the effectiveness of our approach on both synthetic data and real-world B2B marketing data. All the experiments are performed on a Window 7 system with 2 CPUs (Intel i5 2.5GHz) and 8G RAM.

3.3.1 The Experimental Setup

Synthetic Data: As aforementioned, we have simulated a small two-dimensional data set with two classes represented by red and blue color, respectively, as shown in Figure 3.2(a). Some data characteristics of the synthetic data are summarized in Table 3.2.

Real-world B2B Marketing Data: In this study, we obtained a B2B marketing data from a large multinational software company. Specifically, we have two sets of customers interested in different products. One is the network appliance denoted as Product A, and the other one is the desktop visualization software denoted as Product B. With "dormant" customer (no activities for six months) records removed, we have remaining $30,475$ customer records ($8,315$ for Product A and $22,160$ for Product B respectively). Each customer record includes 49 profile attributes and a binary class label to indicate buyer or otherwise.

More specifically, the B2B marketing data set includes demographic attributes, such as industry, company size, and job title information. We also have detailed customer behavior attributes (valued by interaction counts) related to customers' interactions with the company through four major types of marketing campaigns, such as *Event* related campaigns, *Offer* related campaigns, *Product Trial* related campaigns, and *Activity* related campaigns. These behavior attributes reveal meaningful insights about how the prospects behave in specific campaign activities and show their preferences of the marketing campaigns. The data characteristics are shown in Table 3.2, and more details of the attributes are summarized in Table 3.3.

Table 3.2: Synthetic and Two Real-world B2B Data Sets.

Data	Size	Positive Class	Negative Class
Synthetic	300	193	107
Product A	8,315	1,680	6,635
Product B	22,160	4,232	17,828

Benchmark Methods: We compare our approaches with other benchmark methods using two base classifiers (LR and SVM), and we summarize all the methods as follows, where the last two methods (KC and PC) are proposed in this work.

1. Single Classification (LR and SVM): A single general classification approach, not segment-wise classification.

2. Segment-wise Classification using K-Means clustering (SW_{LR} and SW_{SVM}): A step-by-step approach to first use K-Means clustering to do segmentation and then develop classification models for each segment.

3. K-Classifiers Segmentation Approach (KC_{LR} and KC_{SVM}): An integrated approach to use K-Classifiers Segmentation Algorithm to find out both meaningful segmentation and optimized classification models.

4. Profile Consistent Approach (PC_{LR}, PC_{SVM}): An integrated approach to use Profile Consistent Algorithm to ensure profile consistency.

Table 3.3: Demographic Variables and Behavior Variables.

Demographics	**Values**
Company Size	Small Business, Enterprise, Unknown
Industry	Heavy Hitters, Potentials
Job Title	IT Staff, IT Manager, Executive, Researcher, Non-IT, Unknown
Behaviors	**Values**
Event	Corporate Event, Trade Show, Conference, Webinar, Seminar, Technology Preview
Offer	Official Website, Direct Mail, Email, Call Center, Search Engine, Web Advertising (third party), Social Media
Product	Product Download, Product Free Trial, Product Renewal, Product Activation, Product Training
Activity	Subscribe, Unsubscribe, Active, Inactive

3.3.2 Comparison of Targeting Performances

We apply ten-fold cross validation to evaluate the targeting/classification performance with six widely used metrics: *Accuracy*, *Precision*, *Recall*, *F-measure*, *Average Precision Score* (AP), and *Area Under Curve* (AUC) [54]. We report the average performances of the cross-validation in Table 3.4 for both the synthetic data and the real-world B2B marketing data. From Table 3.4, we have the following observations:

- First, we clearly see that our PC approach outperforms other baselines under different metrics, which demonstrates the effectiveness of our framework for buyer targeting optimization. In general, PC has a much better classification performance than KC due to the profile-consistency property of PC.

- Second, PC outperforms other baselines with a more significant margin on synthetic data than on the B2B marketing data. The class ratio of the synthetic data is more balanced, and PC shows a great advantage in terms of all six measures. While for the B2B marketing data, PC achieves better results in terms of *Precision*, *Recall*, *F-measure* and AP. This may be due to that the B2B marketing data is more sparse and with relatively high imbalanced class ratio.

- Third, with utilization of either logistic regression or SVM base classifier, our approach works well and stable on all the data sets. This observation well affirms the advantage of our approach that it is very flexible and can adopt any loss functions.

- Lastly, in some cases, the segment-wise classification approach may have similar performance or even slightly outperform our approach. However, for these two classic and challenging marketing problems, our approach is an innovative attempt for enhancing the marketing performance in a new way and can get improvement from baselines. Also, it is worth noting that the greater performance improvement from KC to PC approach than from SW to PC.

3.3.3 Comparison of the Clustering Performances

In addition to targeting performance, we also compare the clustering performance of our approaches to the benchmark methods. Since there is no external clustering labels as true labels, so we measure the goodness of the clustering results using some internal clustering validation measures based on the compactness and separation.

Clustering Evaluation Metrics: $Calinski - Harabasz$ (CH) index, I index and $Silhouette$ index are used are the evaluation measures [44]. The formulas to calculate the metrics are as follows:

$$CH : \frac{\sum_i n_i d^2(c_i, c)/(NC-1)}{\sum_i \sum_{x \in S_i} d^2(x, c_i)/(n - NC)},$$

$$I : (\frac{1}{NC} \times \frac{\sum_{x \in D} d(x, c)}{\sum_i \sum_{x \in S_i} d(x, c_i) \times \max_{i,j} d(c_i, c_j)})^P,$$

$$Silhouette(S) : \frac{1}{NC} \sum_i \{\frac{1}{n_i} \sum_{x \in S_i} \frac{b(x) - a(x)}{max[b(x), a(x)]}\},$$

where D: data set; n: number of objects in D; c: center of D; P: attributes number of D; NC: number of clusters; S_i: the i-th cluster; n_i: number of objects in S_i; c_i: center of S_i; $d(x, y)$: distance between x and y; $a(x) = \frac{1}{n_i - 1} \sum_{y \in C_i, y \neq x} d(x, y)$; $b(x) = \min_{j, j \neq i}[\frac{1}{n_j} \sum_{y \in C_j} d(x, y)]$.

The CH index validates the cluster performance based on the average between- and within-cluster sum of squares. Index I (I) measures both separation and compactness in terms of the maximum distance between cluster centers and the sum of distances between objects and their cluster center respectively. The *Silhouette* index measures the clustering validity based on the pairwise difference of between- and within-cluster distances. In addition, the larger the values of these three metrics, the better the clustering results.

Except for the single classification approaches, we compare the clustering performances of our proposed approach (KC and PC approaches) to the Segment-wise (SW) approach using K-Means clustering with the aforementioned three metrics. As the experiment results show in Table 3.5, in general, all methods can achieve better clustering results on synthetic data set than real-world B2B data sets. This is because the real-world data sets are more complicated with higher dimensionality than the synthetic data set. Moreover, we can see that our KC_{SVM} approach performs slightly better than Segment-wise approach based on K-means clustering on the synthetic data set. While for the two real-world data sets, KC_{LR} and PC_{SVM} can achieve significant higher values of three metrics. This observation well demonstrate that the our approaches can also achieve great clustering results, and the unified clustering and classification approach can mutually benefit each other.

3.3.4 Decision Oriented Segmentation Analysis

Our approach also provides deep insights in addition to the improved targeting performances. Through the integrated process, distinct classification models for each segment are developed to capture various segment-wise buying decision preferences. Particularly, the attribute coefficients in the classification models can reflect the significance of the impact on the buying decision. Taking Product A for example, we plot the absolute values of the attribute coefficients for five segments as shown in Figure 3.4. The darker the cell color, the larger the value of the coefficient.

As can be seen, each segment has different and diversified sets of significant features, which reveals the different characteristics and buying preferences of the specific segment. Moreover, the set of important features of segment 1 to 5 move from left to right. In general, for customers interested in product A, most important features are among the *Offer* related campaigns and *Event* related campaigns. In contrast, the *Product* related and *Activity* related campaigns are less influential.

To further understand the decision preference of the customers, we list the top influential variables for several segments in Table 3.6. As can be seen, each segment shows different significant variables. 1) For example, for Segment 1 of Product A, four out of five features are the *Offer* related campaigns, namely, website advertising, the company official website and the search engine advertising. Thus, we may summarize the buying preferences of Segment 1 as *Offer Campaign Oriented* segment. 2) It is worth noting that Segment 3 is defined as *Job Title Oriented* due to the job title attributes, such as "Non IT" and "Researcher", which indicate that in this segment customers with these job titles have a more apparent buying decision pattern.

In addition, the differences of the decision preferences exist not only among segments, but also between the two products. As we mentioned before, the buying preferences of Product A's segments mostly focus on the *Offer* related campaigns and *Event* related campaigns. In contrast, Product B have various kinds of buying preference patterns. For example, Segment 1 seems to be *Product Campaign Oriented* due to the majority of the features are product campaigns such as free trial, the total number of product campaign participated in, product activation, product training. Moreover, Segment 3 demonstrates the characteristic of *Activity Campaign Oriented*. These mentioned significant variables in Product B are quite different from those of segments in Product A, which indicates that the customers of these two products behave differently. The above examples show that our approach can grasp the diversified decision preferences of different segments. In summary, the results can help the marketing managers to optimize investment on more efficacious decision preference oriented campaign strategies.

3.3.5 Parameter Sensitivity

In our algorithm, there are two parameters, K and M, which represents the number of segments and the number of subregions, respectively. We fix any one of them and investigate the sensitivity of the other one in turns. For the sake of simplicity, we only show the parameter tuning experiment results of Product B data.

First, we show how to decide the optimal number of clusters K. As shown in Figure 3.5a, we plot the classification performances with increasing number of clusters by fixing $M = 150$. As can be seen, the performances in terms of different measures ($F - measure$ and AUC) vary significantly with the different number of clusters. It is worthy noting that the performances might not increase with either smaller or larger number of clusters. The reason is that, on the one hand, a smaller number of clusters may not be enough to capture the real natural clusters, on the other hand, a larger number of clusters may break the true shape of the natural clutters. Based on Figure 3.5a, we see that K in range of $\{3, 4, 5, 6\}$ have similar performances but $K = 3$ is a feasible choice which we achieve the optimal performance consistently in terms of these two measures. Therefore we choose $K = 3$ for this data set.

Second, the other parameter M, which controls the number of sub-regions using in the Algorithm 2 can also be chosen according to the classification performance. Intuitively, the sub-regions are formed to capture the closeness between the customers. Similarly to the parameter K, either a small or large number of M may not represent the closeness well or even make the sub-regions too trivial. To choose the optimal M, Figure 3.5a shows the classification performances with increasing value of M with fixed $K = 3$, where $M = 150$ gives the consistently optimal results in terms of F-measure and AUC.

3.4 Related Work

Customer segmentation is one of the principal components of CRM since it helps to gain a deep understanding of customers' needs and characteristics [66, 31, 52, 64]. Many data mining techniques are gaining popularity in the market segmentation, such as CHAID decision tree [11, 29, 38], logistic regression [49], neural network analysis [65, 7], and K-Means clustering analysis [26]. In contrast, we formulate an optimized problem with advantages of providing a concrete segmentation focusing on buying decision preferences.

Another key problem in CRM is buyer targeting, that is, to identify the prospects that are most likely to become customers or most valuable to the company. Many database marketers are applying intelligent data mining tools to solve the problem, such as in [6] the authors focused on classification of online customers based on their online website behaviors, and Kim et al. [34] applied neural networks guided by genetic algorithms to target households. Comparing to the previous work that focus on providing a general predicting model for the total customer base, our approach provides an optimized segment-wise approach which can offer more customized and tailored strategies for each segment to improve the customer conversion rate.

Furthermore, the idea of using of segmentation to help build segment-wise prediction models has been recognized by many researchers. Several previous work [18, 3, 2, 58, 25] combined the segmentation and prediction together, and applied on the different business scenarios.

However, the problem for the existing works is that the combination of these two tasks is in a simple step-by-step way, which is difficult to theoretically guarantee the improvement for classification performance. In contrast, our work is distinguished by our development of a joint optimized classification framework, in which the two tasks are unified in a mutually supportive way.

In terms of general-purpose clustering research, this work is related to [41]. As shown, a specific cluster center should be computed for a given distance/loss metric used in the clustering process. In our case, the cluster center is modeled as a classifier to improved the overall classification performance. In other words, our work in this paper is an attempt to unify the supervised and the unsupervised learning methods.

3.5 Summary

Now we answer the question asked in the beginning of this chapter: We can indeed optimally integrate the two essential marketing tasks, customer segmentation and buyer targeting, so that the customers are grouped into segments where the promising buyers can be most easily identified. In our approach, the two tasks are performed simultaneously in a unified optimization framework which combines the clustering and classification objectives. To solve the optimization problem, we developed an iterative *K-Classifiers Segmentation* algorithm, where the customer segments are formed with customers' buyer targeting models. Moreover, we showed that the segmentation results can also be consistent with the features on customer profiles. Finally, we applied our approach on both synthetic data and real-world Business-to-Business (B2B) marketing scenarios. Extensive experiments clearly validated the effectiveness of the proposed approach and its improvements in comparison with alternative methods. In addition to targeting (classification) accuracies, we showed that our approach can provide interpretable customer segmentation solutions and reveals new marketing insights.

Algorithm 1 K-Classifiers Segmentation Algorithm.

Input: X, Y, K, $loss$.

Output: S, C.

for $n = 1, \cdots, N$ **do**
  $\ell_n \leftarrow rand\{1, \cdots, K\}$.
end for
repeat
  #Update step:
  for $k = 1, \cdots, K$ **do**
    Learn C_k based on $\{x_n, y_n | n \in S_k\}$.
  end for
  #Assignment step:
  for $n = 1, \cdots, N$ **do**
    $\ell_n \leftarrow \arg\min_k \text{loss}(x_n, y_n | C_k)$.
  end for
until Convergence.

Algorithm 2 Profile-Consistent Algorithm.

Input: X, Y, K, M, $loss$.

Output: S, C.

```
Randomly select M seed points to construct the Voronoi decomposition T_1, ..., T_M.
for m = 1, ..., M do
    ℓ_m^t ← rand{1, ..., K}.
end for
repeat
    #Update step:
    for k = 1, ..., K do
        Learn C_k based on {x_n, y_n | n ∈ S_k}.
    end for
    #Assignment step:
    for m = 1, ..., M do
        ℓ_m^t ← argmin_k Σ_{n∈T_m} loss(x_n, y_n | C_k).
    end for
until Convergence.
```

Randomly select M seed points to construct the Voronoi decomposition $T_1, \cdots, T_M$.

for $m = 1, \cdots, M$ **do**

  $\ell_m^t \leftarrow rand\{1, \cdots, K\}$.

end for

repeat

  #Update step:

  for $k = 1, \cdots, K$ **do**

    Learn C_k based on $\{x_n, y_n | n \in S_k\}$.

  end for

  #Assignment step:

  for $m = 1, \cdots, M$ **do**

    $\ell_m^t \leftarrow \arg\min_k \sum_{n \in T_m} \text{loss}(x_n, y_n | C_k)$.

  end for

until Convergence.

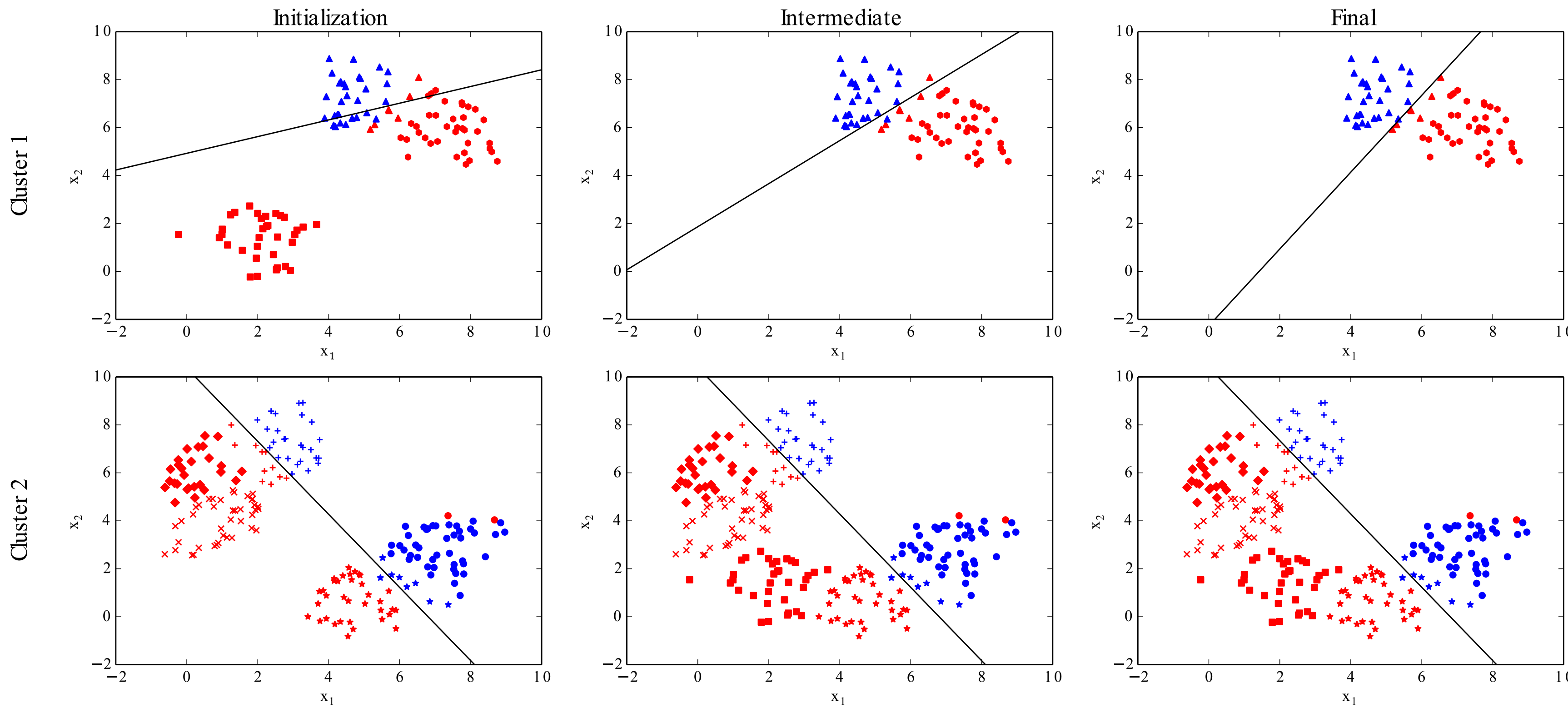


Figure 3.3: An illustration of Profile-Consistent K-Classifiers Segmentation Algorithm. Initialization (left): $M = 8$ sub-regions with different shapes are randomly assigned to $K = 2$ clusters (top and bottom). The classification boundary in each cluster is also plotted. Intermediate results (middle): Sub-regions are re-allocated to cluster 1 (top) or cluster 2 (bottom), according to their respective loss. Final result (right): The algorithm identified two segments both of which are linearly separable.

Table 3.4: The comparisons of targeting performances. For all methods, all parameters (if any) are empirically selected through cross-validation.

		Accuracy	*Precision*	*Recall*	*F − measure*	*AP*	*AUC*
Synthetic (K=2, M=8)	LR	0.6466 ± 0.07	0.7295 ± 0.05	0.7492 ± 0.14	0.7325 ± 0.07	0.8226 ± 0.03	0.5964 ± 0.08
	SW_{LR}	0.8800 ± 0.06	0.9441 ± 0.05	0.8744 ± 0.08	0.9052 ± 0.05	0.9510 ± 0.03	0.8831 ± 0.06
	KC_{LR}	0.5296 ± 0.16	0.6861 ± 0.15	0.5279 ± 0.16	0.5918 ± 0.15	0.7636 ± 0.10	0.5299 ± 0.17
	$\mathbf{PC_{LR}}$	**0.9499** ± 0.05	**0.9651** ± 0.03	**0.9600** ± 0.05	**0.9618** ± 0.04	**0.9759** ± 0.02	**0.9450** ± 0.05
	SVM	0.7039 ± 0.08	0.8272 ± 0.14	0.7692 ± 0.18	0.7686 ± 0.07	0.8751 ± 0.05	0.6746 ± 0.13
	SW_{SVM}	0.7935 ± 0.10	0.9018 ± 0.12	0.8092 ± 0.16	0.8345 ± 0.10	0.9188 ± 0.05	0.7846 ± 0.14
	KC_{SVM}	0.5873 ± 0.10	0.8630 ± 0.13	0.4834 ± 0.17	0.5897 ± 0.14	0.8444 ± 0.05	0.6367 ± 0.09
	$\mathbf{PC_{SVM}}$	**0.9268** ± 0.06	**0.9688** ± 0.03	**0.9200** ± 0.07	**0.9422** ± 0.05	**0.9710** ± 0.02	**0.9300** ± 0.05
		Accuracy	*Precision*	*Recall*	*F − measure*	*AP*	*AUC*
Prod-A (K=5, M=60)	LR	0.8922 ± 0.01	0.7737 ± 0.02	0.6536 ± 0.03	0.7081 ± 0.03	0.7483 ± 0.02	0.8028 ± 0.01
	SW_{LR}	0.8942 ± 0.01	0.7798 ± 0.03	0.6554 ± 0.03	0.7132 ± 0.02	0.7530 ± 0.02	**0.8055** ± 0.01
	KC_{LR}	0.5547 ± 0.13	0.2774 ± 0.20	0.5406 ± 0.06	0.3471 ± 0.14	0.4549 ± 0.12	0.5494 ± 0.10
	$\mathbf{PC_{LR}}$	**0.8954** ± 0.01	**0.7875** ± 0.05	**0.6578** ± 0.04	**0.7146** ± 0.03	**0.7559** ± 0.03	0.8054 ± 0.02
	SVM	0.8117 ± 0.05	0.5482 ± 0.13	0.5495 ± 0.09	0.5414 ± 0.09	0.5939 ± 0.08	0.7133 ± 0.05
	SW_{SVM}	0.8255 ± 0.04	0.5813 ± 0.12	**0.5875** ± 0.10	0.5738 ± 0.08	0.6257 ± 0.07	**0.7363** ± 0.04
	KC_{SVM}	0.5399 ± 0.03	0.2222 ± 0.02	0.5194 ± 0.07	0.3108 ± 0.03	0.4188 ± 0.04	0.5322 ± 0.03
	$\mathbf{PC_{SVM}}$	**0.8504** ± 0.01	**0.6635** ± 0.04	0.5165 ± 0.09	**0.5766** ± 0.06	**0.6383** ± 0.04	0.7252 ± 0.04
		Accuracy	*Precision*	*Recall*	*F − measure*	*AP*	*AUC*
Prod-B (K=3, M=150)	LR	0.9277 ± 0.01	0.8555 ± 0.01	0.7682 ± 0.05	0.8087 ± 0.03	0.8350 ± 0.02	0.8679 ± 0.02
	SW_{LR}	0.9281 ± 0.01	**0.8557** ± 0.01	0.7705 ± 0.05	0.8102 ± 0.03	0.8360 ± 0.02	0.8690 ± 0.02
	KC_{LR}	0.6053 ± 0.12	0.3120 ± 0.19	0.5737 ± 0.04	0.3864 ± 0.12	0.4855 ± 0.11	0.5935 ± 0.08
	$\mathbf{PC_{LR}}$	**0.9292** ± 0.00	0.8556 ± 0.01	**0.7771** ± 0.02	**0.8143** ± 0.01	**0.8386** ± 0.01	**0.8721** ± 0.01
	SVM	0.8551 ± 0.02	0.6313 ± 0.07	**0.7042** ± 0.07	0.6610 ± 0.04	0.6973 ± 0.03	0.7985 ± 0.03
	SW_{SVM}	0.8534 ± 0.03	0.6285 ± 0.07	0.6830 ± 0.09	0.6500 ± 0.06	0.6875 ± 0.05	0.7895 ± 0.04
	KC_{SVM}	0.5346 ± 0.07	0.2259 ± 0.03	0.5347 ± 0.09	0.3157 ± 0.04	0.4268 ± 0.05	0.5346 ± 0.05
	$\mathbf{PC_{SVM}}$	**0.8961** ± 0.01	**0.7736** ± 0.04	0.6871 ± 0.09	**0.7230** ± 0.05	**0.7616** ± 0.03	**0.8177** ± 0.04

Table 3.5: The comparisons of clustering performances.

		$K-Means$	KC_{LR}	PC_{LR}	KC_{SVM}	PC_{SVM}
Syn	$Silhouette$	0.9178 ± 0.01	0.9270 ± 0.01	0.9226 ± 0.01	$\mathbf{0.9278 \pm 0.01}$	0.9277 ± 0.01
	CH	1021.87 ± 48.1	1099.95 ± 62.3	1037.49 ± 90.5	$\mathbf{1105.113 \pm 56.3}$	1100.80 ± 54.9
	I	4373 ± 349	4661 ± 377	4429 ± 451	$\mathbf{4839 \pm 312}$	4693 ± 350
Prod-A	$Silhouette$	0.1971 ± 0.03	$\mathbf{0.2228 \pm 0.01}$	0.2206 ± 0.01	0.2036 ± 0.00	0.2088 ± 0.01
	CH	564.13 ± 9.61	$\mathbf{572.68 \pm 8.05}$	560.83 ± 4.08	568.92 ± 6.25	570.38 ± 5.73
	I	$2.16E+106$	$\mathbf{3.76E+113}$	$2.19E+112$	$1.43E+107$	$5.51E+106$
Prod-B	$Silhouette$	0.1854 ± 0.01	0.2122 ± 0.03	0.1924 ± 0.02	0.2010 ± 0.01	$\mathbf{0.2132 \pm 0.02}$
	CH	1322.96 ± 25.8	1351.2 ± 31.5	1343.18 ± 23.62	1372.06 ± 21.75	$\mathbf{1384.78 \pm 27.48}$
	I	$3.80E+102$	$2.81E+109$	$1.03E+107$	$3.09E+108$	$\mathbf{1.51E+111}$

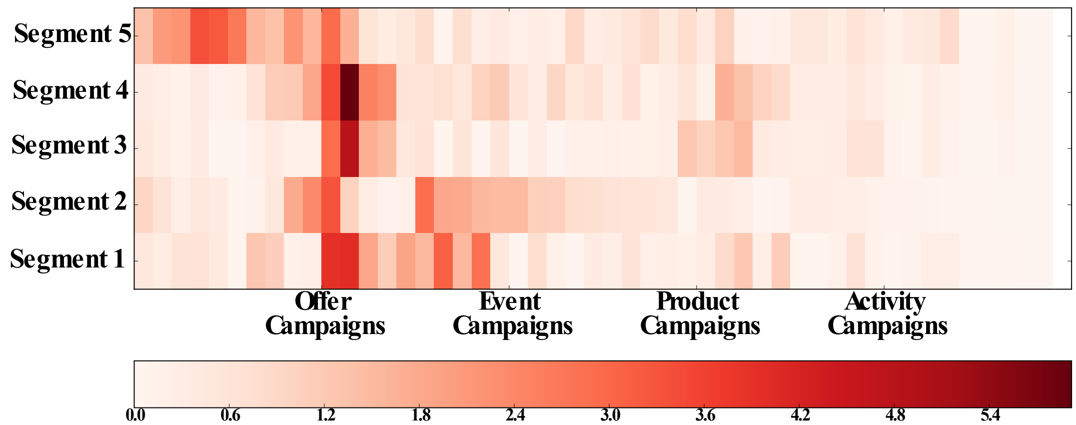
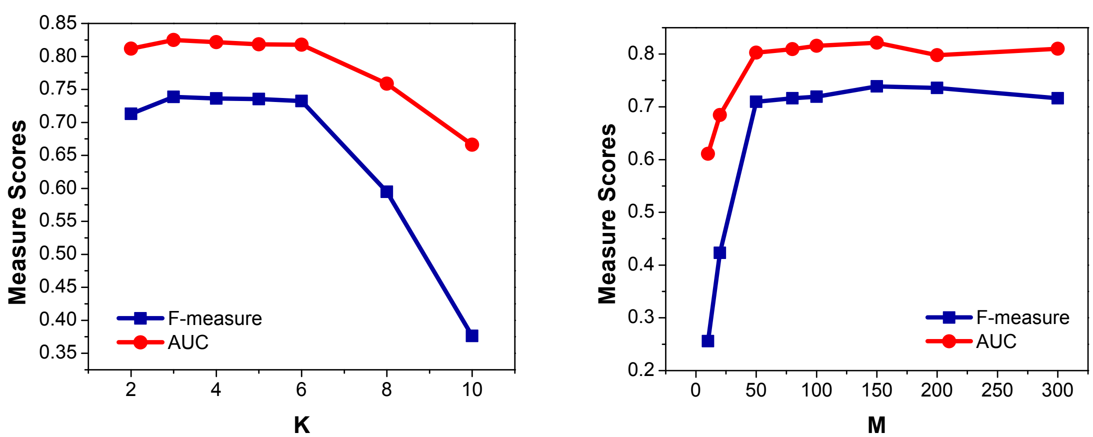


Figure 3.4: A heat map of the targeting coefficients of profile variables for Product A.



(a) (b)

Figure 3.5: (a): Impacts of The Number of Clusters on Model Performances. (b): Impacts of The Number of Sub-Regions on Model Performances.

Table 3.6: The top 5 most significant variables per segment for Product A & B.

Segment	Significant Variables	Property
	Product A	
S1	**Offer-WebAds**, **Offer-OfficialSite**, **Offer-SearchEngine**, **Offernumtotal**, Evnt-Webinars	Offer Campaign Oriented
S2	**Evntnumtotal**, **Evnt-Corpevents**, **Evnt-Webinars**, **Evnt-Conferences**, Activitynumtotal	Event Campaign Oriented
S3	Evntnumtotal, **Non-IT**, Activitynumtotal, **Researcher**, Evnt-Webinars	Job Title Oriented
S4	**Offer-SocialMedia**,**Offer-CallCenter**, Executive, **Offer-Email**, **Offer-Directmail**	Offer Campaign Oriented
S5	**Evnt-Tradeshow**, **Evnt-Seminars**, **Evnt-Webinars**, Offer-Email, **Evnt-TechPrev**	Event Campaign Oriented
	Product B	
S1	**Prod-Trialfree**, **Prodnumtotal**, **Prod-Activation**, Evnt-numtotal, **Prod-Training**	Product Campaign Oriented
S2	**Evnt-Corpevents**, **Evnt-Webinars**, Prod-Trialfree, **Evnt-Seminars**, **Evnt-Webinars**	Event Campaign Oriented
S3	**Activ-Subscribe**, **Activein3m**, **Activitynumtotal**, Offer-SearchEngine, **Activ-Unsubscribe**	Activity Campaign Oriented

CHAPTER 4

MULTI-FOCAL LEAD SCORING FOR PREDICTIVE MODELING IN BUSINESS-TO-BUSINESS MARKET

An important aspect of customer acquisition in business-to-business (B2B) marketing is lead prospecting which guides the business to select the right leads to pursue. While the current practice is often to use arbitrary rules to select leads with ad-hoc basis or pure managerial intuitions, a more systematic and data-driven approach is needed to improve the quality of lead selection by quantifying a specific lead's tendency to become a customer. To this end, in this chapter, we introduce a predictive lead scoring model which can help sales representatives to identify prospective leads from a large pool of candidates in a B2B environment. Existing studies along this line have a focus on predicting "propensity to buy", yet limited efforts have been made on exploring discrepancy between lead segments. In response, we provide a multi-focal lead scoring framework which can improve the performance of predictive lead scoring. Specifically, in this framework, leads are first divided into several focal groups (segments) based on their characteristic attributes (features) and marketing workflows. Then, a logistic regression scoring model is learned for each segment with multi-task learning (MTL) technique, a machine learning approach to jointly constructing predictive models for multiple focal groups. Indeed, the key of multi-focal learning in this study is to allow predictive modeling in each segment consisting of leads with similar characteristics rather than modeling the whole population of leads with varying characteristics. However, independent modeling at focal level would be problematic for segments with few representative samples. We use the MTL framework to address this problem by exploiting commonalities shared by focal groups and automatically balancing between unification of all groups and individualization of each group. Finally, empirical findings derived from real-world B2B marketing data demonstrate that different segments may have absolutely different conversion rates and leads in the same segment tend to have similar responses to a specific marketing campaign. Therefore, such a multi-focal tailored lead scoring model gives a better insight into factors influencing the conversion of leads.

4.1 Introduction

In the big data era, marketers rely on data analytics to ensure that they are equipped with the best insights to stay tuned in today's dynamic business environment. In Business-to-Business (B2B) marketing, successful marketers can collect plenty of information about potential customers, which are also known as *leads*. The information usually consists of demographic attributes (e.g., job title, education background), firmographic attributes (e.g., company industry and size), and behavioral records (e.g., online footprints from an email click, webpage visits, white paper downloads, or other marketing activity touch-points). However, the vast amount of information seems overwhelming to B2B marketers [68]. It is practically infeasible for sales team to contact everyone who has expressed any level of interests. Thus, automatically selecting promising leads from a large pool of customers is an essential B2B marketing tool. With high predictive power in identifying the right leads to pursue, a *lead scoring* tool can significantly improve the effectiveness of lead generation and the overall marketing performance of enterprises.

Indeed, data-driven lead scoring has been a critical component of B2B marketing strategy. A lead scoring model can score leads with respect to their readiness or maturity for sales. Marketers usually prioritize leads with higher scores as more promising candidates for sales, thus leads with scores that meet or exceed a threshold are passed along to the sales team for marketing engagement. The scores can also be transformed to ranks such as "hot", "warm" or "cold" with multiple thresholds, and sales team can the use different engagement strategies to interact with customers from different ranks.

Therefore, a well-planned, carefully-designed, and optimized lead scoring model is a key contributor to marketing productivity, sales engagement, and the rate at which inbound leads convert to sales opportunities and business deals. However, existing empirical approaches of lead scoring are often not optimized. For example, by fitting historical sales records and attributes of customers with statistical models (e.g., logistic regression), a lead scoring model can be a simple summation of scores of all attributes of customers.

Due to the existence of heterogeneity of lead behaviors, building just one scoring model for all records may not be appropriate. To alleviate this problem, customer segmentation has also been utilized for developing better leading scoring models. The idea is to first split the overall population of customers into several homogeneous focal segments and then build a lead scoring model specifically for each segment. By doing so, [20, 51, 3] revealed that segmentation-based lead scoring models can be much more effective in terms of predictive power.

Nonetheless, there are still major issues with the existing approaches. In particular, customers from different segments may have different attributes and behave differently, they still share commonalities given that they are all leads of the same marketer or in the same market environment. However, aforementioned independent focal-level lead scoring approach completely ignores any relationships among different segments by not exploring commonalities among leads from different segments. Indeed, it's a nontrivial endeavor to quantify and utilize global commonalities among focal customer segments for improved lead scoring models and marketing performance.

In this chapter, we develop a multi-focal leading scoring framework which can automatically balance between unification of all customer groups and individualization of each group. The key difference between our *multi-focal leading scoring* and the aforementioned *independent scoring models at focal levels* is that, our framework can jointly build predictive models for all groups by exploring their structural connections. By doing so, we address several important challenges for predictive modeling in B2B marketing:

- Customers from different market segments have different demographic and firmographic attributes, and may react differently toward the same marketing engagement strategy. As a consequence, customer attributes should have different levels of significance for different customer segments in the lead scoring models. Our framework can take into account differences among market segments so that the lead scoring performs better than one global scoring model for all customer segments. The modeling results can answer questions such as: how likely a prospect will be converted to sales opportunity; what are the top predictors; how strong each predictor is in different segments.

- Our framework can quantify and utilize structural connections among market segments for improved lead scoring performance. In fact, effective lead scoring require both sufficient sample size and conversion rate, where the sample size is the number of customers fitting the model and the conversion rate is the ratio between converted leads and non-converted leads. The customer segmentation strategy may reduce both sample size and conversion rate for some segments, resulting in the imbalanced learning problem (also known as rare class problem) [69]. By exploring connections among customer groups, multiple lead scoring models can reinforce each other and effectively address the imbalanced learning problem. In other words, segments with good modeling performance can help other segments while maintaining relative independence.

Our multi-focal lead scoring framework consists of two phases. The first phase is to form focal groups, where each focal group is one unique segment of leads on market. To this end, we exploit CHAID [32], one popular customer segmentation tool in marketing research, to form the lead segmentation. As a decision tree technique, CHAID uses adjusted significance testing (Bonferroni testing) to construct multi-split trees of customers with both categorical and continuous variables, and the segmentation result is easy to interpret represented with an intuitive tree structure. Of note, we want to emphasize that, though we use CHAID in this work, our framework is flexible to use alternative clustering approaches to discover the multi-focal groups.

In the second phase, we jointly build lead scoring models for multiple focal groups of customers. We will use Multi-Task Learning (MTL), a machine learning approach to quantify and utilize structural connections among market segments. Here, building the lead scoring model for each customer segment is a *learning task* in the MTL approach. The key benefit of MTL is its ability to improve the overall learning performances by transferring knowledge among the learning tasks.

In summary, this study contributes to the understanding of lead prospecting in several ways. First of all, we propose a multi-focal lead scoring framework that can significantly improve the quality of lead prospecting for B2B marketing applications. In particular, this multi-focal framework allows marketers to gain better marketing insights, such as what types of leads or lead characteristics matter most. Second, our findings provide empirical evidence about how leads interact with marketing nurturing campaigns and shed new light on the drivers of lead responses to a firm's marketing campaigns. Finally, the proposed framework has a potential to be generalized to other marketing and business scenarios, such as brand promotion, product cross-selling or up-selling, and targeted advertising. To the best of our knowledge, how to score the readiness of the prospects using data-driven techniques and then deploy proactive marketing campaign management is still under-explored, and our work is the first deployment of advanced predictive learning in address these questions.

Overview. The rest of this chapter is organized as follows. In Section 4.2, we discuss related work on predictive modeling in marketing related research areas. In Section 4.3, we discuss how to do processing B2B marketing data, including the data-mining techniques used to identify different multi-focal groups (segments), and the methods used to identify important characteristics of highly potential prospects. Section 4.4 gives an overview of the multi-focal lead scoring approach. In Section 4.5, we show empirical results. Finally, Section 4.7 gives the conclusions, marketing implications, and future research directions.

4.2 Related Work

The related literature can be grouped in two categories: customer segmentation for predictive marketing models and multi-task learning (MTL) for marketing applications.

4.2.1 Customer Segmentation For Predictive Marketing Models

First, customer segmentation is one of the most important components of Customer Relationship Management (CRM) since it helps to gain a deep understanding of customers' needs and characteristics [**?** 52]. The objective of customer segmentation is to partition a heterogeneous group of customers into several identifiable and internally homogeneous subgroups based on both demographical characteristics and purchasing behavior of customers. After segmenting customers, firms can offer differentiated marketing strategies to targeted customer groups to improve the marketing ROI (Return on Investment). Therefore, customer segmentation is recognized as a fundamental aspect of customer-centric marketing paradigm [66, 64, 31].

Well-known approaches to customer segmentation include both conventional models such as RFM (Recency, Frequency, and Monetary) [5] and FRAC (Frequency, Recency, Amount (of Money) and Category (of Product)) [33], as well as recently developed data mining techniques. With the increasing volume of big marketing data, data mining methods are gaining popularity in the application of market segmentation, such as clustering analysis [56, 59, 21], Neural Networks [65, 7, 16], and especially Decision Trees [32, 52]. Using one specific decision tree model, CHAID (CHi-square Automatic Interaction Detection), Haughton & Oulabi [29] studied the performance of direct marketing based on customer segmentation; Bult & Wansbeek [11] devised a profit maximization approach to select customer segments; and Levin & Zahavi [38] studied CHAID using logistic regression model as a benchmark for comparative analysis of customers.

Customer segmentation can improve the effectiveness of predictive models in sales forecasts or prospect acquisition. The reason is that, a general and aggregated predictive model for all customers is less effective and can be misleading given the discrepancies among customer segments [2]. In the literature, several studies have investigated this hypothesis. For instance, Currim [20] revealed that a segment-based model is preferable to an aggregate model, depending on the extent of heterogeneity in the market. Also, people have shown that more accurate predictive results can be obtained by first using specific methods to segment customers [51, 37]. Moreover, in some cases, segmentation-based models are more effectively for improving predictive accuracy in targeted marketing [3]. Finally, Reutterer et al. [58] proposed a dynamic segmentation approach for targeting and customizing direct marketing campaigns, and they demonstrated the procedure in a controlled field experiment. However, the existing approaches follow a simple segmentation-and-prediction procedure and ignore the potential risk of segmentation. Indeed, in B2B marketing, prospect acquisition is usually a rare-class prediction problem; that is, there are far more less converted prospects than the candidate prospects. In other words, there will be short of representative converted prospects in the whole data. After segmentation, this issue will become more salient in some segments and the learning performances in these segments will be deteriorated instead of being improved. Therefore, in this work, we develop a multi-focal learning framework to address the new challenges, in particular for B2B marketing.

4.2.2 Multi-task Learning for Marketing Applications

When there are intrinsic relatedness among separate tasks to learn, Multi-task learning (MTL) takes advantage of the relatedness by learning all tasks simultaneously instead of following the traditional single task learning (STL) approach which learns each task independently. Some experimental studies proved the benefits of MTL compared to STL when the tasks are related. Caruana [14] summarized previous works on MTL and presented evidences to show that MTL in backprop nets utilized task relatedness to improve performances with k-nearest neighbor, kernel regression and decision tree algorithms. Also, Bakker & Heskes [4] implemented multi-task learning as a hierarchical Bayesian approach on a well-known clustering school problem. This study showed that the tasks were modeled better through Bayesian multitask learning. There are also theoretical studies on MTL about different assumptions on task relatedness. For example, mean regularized MTL assumes all the tasks are related and the task-specific models come from a common distribution [23].

Generally speaking, our multi-focal learning framework is specially designed for predictive modeling in B2B market. First of all, the customer segmentation is necessary for prioritizing customer handling and marketing interventions according to the importance of each customer. However, predictive modeling in B2B market is essentially a rare-class learning problem. After segmentation, the imbalanced class distribution can be improved in some segments; that is, the percentages of converted prospects in these segments are larger than that of converted prospects in the whole data. As a result, the learning performances in these segments can be naturally improved. In contrast, the imbalanced class distribution can become more skewed in the rest segments; that is, the percentages of converted prospects can be reduced in these segments compared to that of converted prospects in the whole data. In other words, less representative converted samples will be available in these segments, and thus the learning performances will be naturally deteriorated. Instead, in our framework, MTL has been exploited as a complement to enhance multi-focal learning. MTL has the ability in leveraging converted samples from other segments to improve the learning performances in a segment with very few representative converted prospects.

4.3 B2B Marketing Data Description

In this study, we obtained the B2B marketing data from a 500 fortune software company. This real-world B2B marketing data set contains marketing activities of all the prospects from 2011 to 2013. Moreover, this data set has imbalanced class distribution, where the number of observations belonging to the conversion class is significantly lower than those belonging to the not-conversion class. Concretely, the conversion rate is 6%, with $2,989$ conversion cases out of total $49,575$ individuals.

The B2B marketing data set includes 4 demographic variables and 36 behavior variables related to four major types of marketing campaigns. Demographic variables include industry, company size, job title and existing/new accounts. Table 4.1 shows the detailed values of demographic variables.

There are two important data preprocessing steps. First, the values of demographic variables are aggregated into high level categories. For example, for the *Industry* variable, there are more than 20 different categorical values, which lead to very sparse results in the segmentation process. In other words, the segmentation at the original level is too sparse to derive meaningful and useful insights of customers buying patterns. Therefore, we merge industry values into 3 categories according to the level of conversion rates and profits. Second, there are a significant number of missing values in the demographic variables. To deal with this situation, we create an "Unknown" value for all the missing entries. The detailed aggregation rules are listed in the Appendix 1.

As for the behavior variables, they depict the insights about how the prospects interact with specific campaign activities, which can indicate the readiness of prospects to convert. Particularly, behavior variables are associated with four major types of marketing campaign activities, such as events, promotion offer (from different channels), product download, and unsubscribe activities. Most behavior variables are binary attributes, "1" indicating the prospect involved in a specific activity and "0" meaning not involved. In addition, only a few variables are numeric attributes de-

Table 4.1: Demographic variables in the B2B marketing dataset

Variable Name	Value
CompanySize	{Small Business, Enterprise, Unknown}
ExistingorNew	{Existing, New}
Industry	{Heavy Hitters, Potentials, Laggards, Unknown}
JobTitle	{IT Staff, IT Manager, Executive, Researcher, NonIT, Unknown}

scribe the frequency of specific activity. From the Figure 4.1 of the pairwise Pearson correlation of these numeric variables, we find that there is a relative weak positive linear correlation between *OffernumOfficialWebsite* and *Productnumtotal*, *OffernumOfficialWebsite* and *OffernumSEM*,*OffernumSEM* and *OffernumWebads*, and *Evntnumconferences* and *Evntnumcorpevents*. But in general, there is no strong linear correlation among these numeric behavior variables.

Moreover, there are some other variables such as "DayssinceMindate" indicating the number of days since the prospect account was created, "Num of Interaction" for the number of total interactions and "Num of Contacts" for the number of total contacts within different time periods (first 3 months or 6 months). Table 4.2 shows the detailed values of behavior variables. Based on above variables, we then develop a scoring model to identify and qualify those prospects as prospective leads.

Table 4.2: Behavior variables in the B2B marketing dataset

Behavior Category	Specific Activity
Event	{Corporate Event, Tradeshow, Conference, Webinar, Seminar}
Offer	{OfficialWebsite, Direct Mail, Email, Search Engine, Web Ads (Third party) }
Product	{Product Download, Product Free Trial}
Unsubscribe	{Unsubscribe}
Other	{DayssinceMindate, Num of Interaction, Num of Contacts}

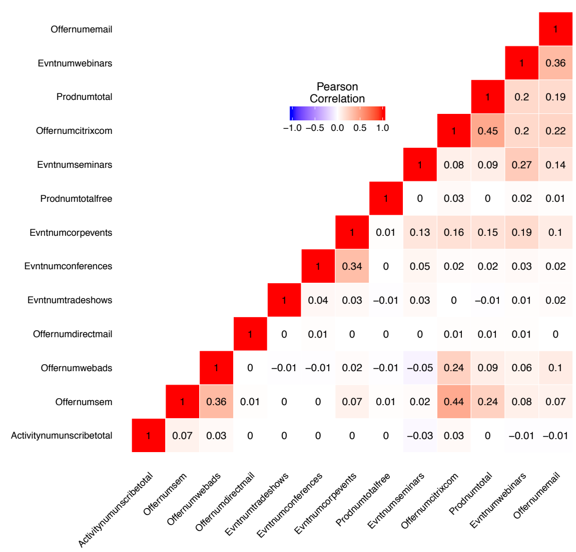


Figure 4.1: The heatmap of the Pearson correlations of numeric variables.

4.4 The Multi-focal Lead Scoring Framework

In this section, we provide a multi-focal lead scoring framework which can produce a lead score for each prospect to represent their readiness and importance based on different multi-focal groups. Figure 4.2 shows the multi-focal lead scoring framework.

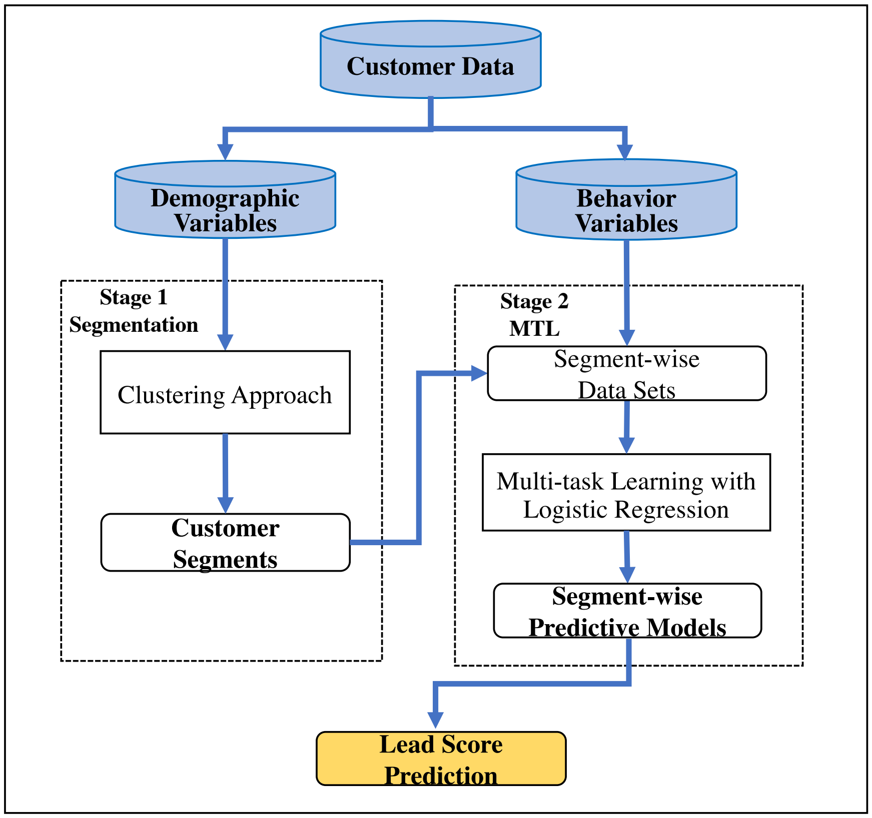
The data set is usually split into two parts: training set and test set. We learn the models on the training set with known customer buying decision (converted or not) and then use the test data set to evaluate the models' prediction performance.

As can be seen, there are two main stages for multi-focal lead scoring. In the first stage, the task is to identify different focal groups (segments) in a way such that training instances in the same group are more similar to each other and instances from different groups are distinct from each other as much as possible. The second stage is to estimate lead scoring models for each segment using multi-task learning (MTL). All the tasks from these segments are related and will be learned simultaneously by MTL. Then, every test sample in each segment will be assigned a lead score based on the corresponding model. Therefore, in this framework, we can prioritize the prospects as qualified leads within each segment, and thus generate a top-k list of the prospective leads for sales representatives to pursue.

4.4.1 The Formation of Multi-focal Groups

The formation of multi-focal groups is important for exploiting the inherent heterogeneity of prospects. Any widely applied clustering approaches can be applied to discover the multi-focal groups. Moreover, we can always form focal groups based on business rules or directly use the given group setting. In this study we employ

Figure 4.2: The Multi-focal Lead Scoring Framework Using Multi-task Learning.



CHAID [29], a popular automatic tree-structured segmentation method in marketing research, for identifying the focal groups of prospects.

The result of the CHAID analysis (using SPSS Statistics Version 22) is shown in Figure 4.3. Here, the segmentation phase is focused on the demographical and firmographical features, while in the second phase, the lead scoring model is built using the behavior variables which indicates the engagement and readiness of the prospects. The combination of both profile and behavior features can provide more enriched information about the prospects, and thus improve the effectiveness of lead scoring models. Hence, when implementing a segmentation scheme, demographic

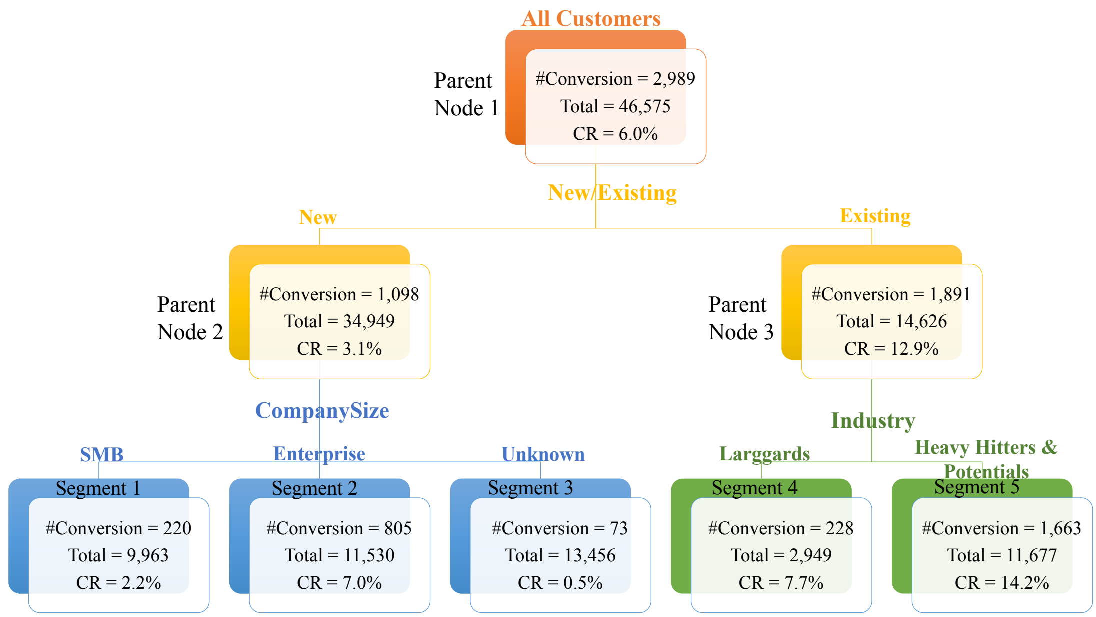


Figure 4.3: The Segmentation Result using CHAID

features can lead to a better delineation of segments by characterizing segments using semantic information [27]. To achieve a small size of terminal nodes, we specified that the level of the CHAID tree is 2. Thus, four variables are used to construct the CHAID tree, namely, *ExistingorNew*, *CompanySize*, *Industry* and *JobTitle*, which are represented using different colors.

Figure 4.3 shows that the most significant splitting variable for segmentation is *ExistingorNew*. For existing prospects, who had business with the company before, the next significant variable is *Industry*. However, for completely new prospects, the next significant variable is *CompanySize*. This result indicates some insights for the existing prospects of the company; that is, *Industry* can segment the prospects better according to chi-square than *CompanySize*. In contrast, for those new prospects, *CompanySize* plays a more important role than *Industry*to distinguish instances in two nodes.

As shown in the tree plot, there are five leaf nodes representing Segment 1 to 5 respectively. We also include the information of size and conversion rate for each segment in the Figure 4.3. The top three segments ranked based on the conversion rates are: Segment 5 (existing prospects in the industry of *heavy hitters* and *potentials*) has the highest conversion rate of 14.2%, Segment 4 (existing prospects in Laggards industry) with 7.7% conversion rate and Segment 2 (new prospects in enterprise) with 7.0% conversion rate. In contrast, Segment 3 has the lowest conversion rate of 0.5% and has a very skewed class distribution. In Segment 3, the directly-built model cannot perform well. Instead, we propose to use MTL in the second phase as a complement to enhance multi-focal learning.

In sum, we may summarize the characteristics for each segment as follows:

- Segment 1: new customers from small and median businesses.
- Segment 2: new customers from enterprise businesses.
- Segment 3: new customers from businesses with unknown company size.
- Segment 4: existing customer from industry type laggards.
- Segment 5: existing customer from both heavy hitter and potential industry.

In addition, we plot histograms to show the distribution of customer interactions with different marketing campaign types across five segments in Figure 4.4 and Figure 4.5. For the product type campaigns, all five segments have consistent preferences with 95% of interactions are *Product Trial* and 5% are *Product free*. As shown in

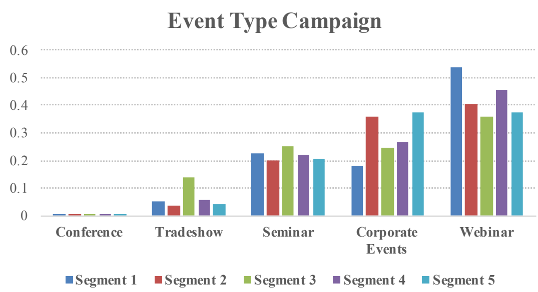


Figure 4.4: Event Type

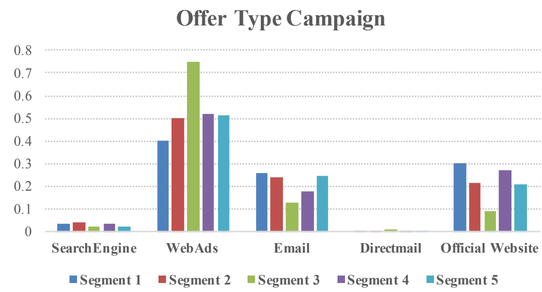


Figure 4.5: Offer Type

the chart, we can clearly see that the proportion of different campaign types is also variant in different segments. This also proves that different segments have distinct preferences and behavior patterns, thus indicating a strong need for clustering customer bases into focal groups. In general, a few campaign types such as *webinar*, *web advertisement*, are always the most popular campaigns in *Event* and *Offer* type respectively. In contrast, *Conference* and *Direct mail* are the least popular campaigns. The different distribution of specific campaign types in different segments shows that the multi-focal property is inherent in segments. Also, we can see the hidden relatedness among these five segments, thus MTL approach can be applied in the second phase.

4.4.2 Multi-task Learning

The second phase of the proposed multi-focal framework is responsible for building a lead scoring model for each segment to rank the prospects based on the propensity to be converted by using their behavior information related to marketing campaigns. The conversion possibility of a prospect can be viewed as the quality or the "score" of the prospect. As a result, the marketers are able to rank all the prospects based on the scores to further prioritize the top prospects as prospective leads, and provide a deep understanding about what kind of campaigns works the best on a specific group of prospects. This is a learning problem which can be solved by predictive modeling techniques. The predictive modeling technique we employ in this study is *Logistic Regression*, which is one of the benchmark "predictive" modeling techniques in marketing research [49, 28, 8].

Based on the segmentation result, there are five predictive models to be learned. However, it may not be optimal to learn a logistic regression model for each segment separately. Indeed, although we believe prospects from different segments will behave in different ways during their corresponding buying processes, the prospects might also share some intrinsic commonalities. Therefore, a data driven procedure is desired to take the discrepancy of the different segments into account and identify the potential common characteristics of their predictive models. To this end, we exploit the Multi-Task Learning (MTL) techniques to simultaneously build the predictive models of all the segments by balancing the model complexity of all these models. In this way, the models of different segments will share a common structure and the information in the data of different segments can be shared during the learning process. These benefits are especially important for some segments containing very few converted samples. In these segments, the separately learning procedure may fail to produce robust results.

In this research, let us consider the following setup. We have T learning tasks associated with our T segments respectively. For the t-th task, we have N_t customers represented by the behavioral attributes $\{x_n^{(t)} | n = 1, \cdots, N_t\}$ and the corresponding

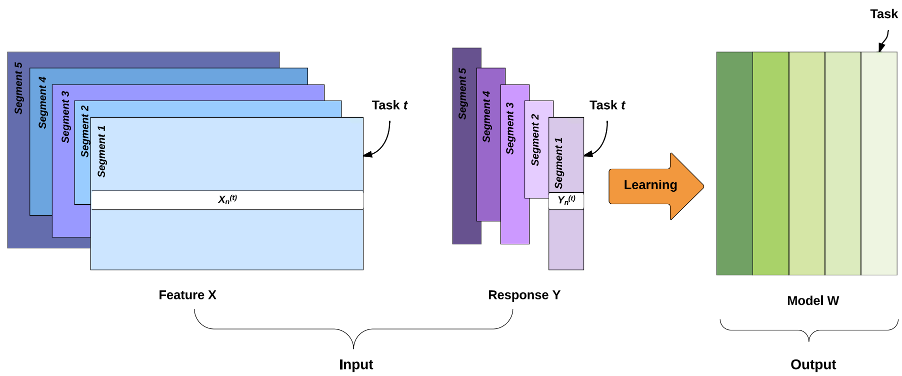


Figure 4.6: An illustration of MTL [79].

conversion label $\{y_n^{(t)}|n = 1, \cdots, N_t\}$. The goal is to learn T functions $f_t(x) = xw^{(t)} + c^{(t)}$ for $t = 1, \cdots, T$ respectively, which can be applied on $x^{(t)}$ to predict $y^{(t)}$. A general approach to this end is to minimize the penalized empirical loss:

$$\min_W \mathcal{L}(W, c) + \Omega(W) \tag{4.1}$$

where $W = (w^{(1)}, w^{(2)}, \cdots, w^{(T)})$, $c = (c^{(1)}, c^{(2)}, \cdots, c^{(T)})$ are the parameters to be estimated, $\mathcal{L}(W, c)$ is the empirical loss on the training samples, and $\Omega(W)$ is a regularization term [23] representing the entire model complexity of all the tasks. When applying logistic regression, we have the loss

$$\mathcal{L}(W, c) = \sum_{t=1}^{T} \sum_{n=1}^{N_t} \log(1 + \exp(-y_n^{(t)}(x_n^{(t)} w^{(t)} + c^{(t)}))). \tag{4.2}$$

An intuitive demonstration of our setup is shown in Figure 4.6.

A general way to formulate the model complexity $\Omega(W)$ is through a graph G encoding the task relatedness. Specifically, G is defined with the tasks as nodes. We have an edge connecting two tasks if and only if they are related. Each edge connecting the i-th and j-th tasks can also be weighted by a positive measurement, denoted by G_{ij}. Then, the regularization can be formalized as

$$\Omega(W) = \lambda \sum_{i=1}^{T} \sum_{j=1}^{T} G_{ij} \|w^{(i)} - w^{(j)}\|^2 = \lambda \mathrm{tr}(WLW'),$$

where $L = D - G$ is the graph Laplacian of G, D is diagonal degree matrix of G such that $D_{ii} = \sum_{j=1}^{T} G_{ij}$, and $tr(\cdot)$ represents the trace. Intuitively, the graph based regularization $\Omega(W)$ imposes constraints on the distance between the model parameters $w^{(i)}$ and $w^{(j)}$ if their corresponding tasks are related, where the constraint degree is controlled by a positive parameter λ. Specifically, a large λ will tend to make the related models to be the same model and the segmentation structure can be ignored. On the other hand, a small λ will tend to make all the tasks unrelated and the problem reduces to solving the T tasks independently. However, the optimal λ is dependent on the actual data characteristics, and it can be achieved by applying the cross validation procedure.

The graph based regularization is quite flexible to encode different assumptions on the task relatedness. For example, [23] proposed the *Mean-induced* regularization by assuming all tasks are related in the way that the models of all tasks are close to their mean, and it can be shown that this is equivalent to apply the graph based regularization with a graph G whose elements are all ones.

More generally, the graph structure can also be provided by domain experts or derived from external knowledge on the tasks. For example, we can derive a graph for multi-task learning based on our hierarchical segmentation structure in Figure 4.3. To be specific, we define two different *Tree-induced* graph G with levelwise granularity and nodewise granularity, respectively.

- MTL-Level Approach

For the levelwise granularity (denoted as MTL-level):

$$G = \sum_{\ell=1}^{L} \phi_\ell G_\ell,$$

where ℓ represents the level of the segmentation tree, and G_ℓ is the adjacency matrix for the ℓ-th level. Moreover, ϕ_ℓ is the weight of ℓ-th level, and we consider that ϕ_ℓ increases when the tree grows deep.

- MTL-Node Approach

For the nodewise granularity (denoted as MTL-node):

$$G = \sum_{p=1}^{P} \phi_p G_p,$$

where p represents the parent nodes of all the leaf nodes, and G_p is the adjacency matrix for the p-th parent node. Similarly, ϕ_p is the weight of p-th parent node, which can represent the behavior similarities among the connected nodes, and we also consider that ϕ_p increases when the tree grows deep.

Let us take the MTL-Node for the segmentation tree in Figure 4.3 as an example, in which there are only two levels with $P = 3$ parent nodes in addition to the leaf nodes. In the first level of the root node, the two children nodes are connected together. However, in the second level, Segments 1, 2, 3 are more related and Segments 4, 5 are more related to each other. Hence, in the first level, we can consider all 5 segments are connected with the weight of ϕ_1. In the second level, segments 1, 2, 3 are linked together with ϕ_2. Also, segments 4, 5 are linked together with ϕ_3. Hence, the G can be shown as the following:

$$G = \begin{pmatrix} \phi_1 & \phi_1 & \phi_1 & \phi_1 & \phi_1 \\ \phi_1 & \phi_1 & \phi_1 & \phi_1 & \phi_1 \\ \phi_1 & \phi_1 & \phi_1 & \phi_1 & \phi_1 \\ \phi_1 & \phi_1 & \phi_1 & \phi_1 & \phi_1 \\ \phi_1 & \phi_1 & \phi_1 & \phi_1 & \phi_1 \end{pmatrix} + \begin{pmatrix} \phi_2 & \phi_2 & \phi_2 & 0 & 0 \\ \phi_2 & \phi_2 & \phi_2 & 0 & 0 \\ \phi_2 & \phi_2 & \phi_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \phi_3 & \phi_3 \\ 0 & 0 & 0 & \phi_3 & \phi_3 \end{pmatrix}$$

We will examine the performances of both the *Mean-induced* regularization and the *Tree-induced* regularization in Section 4.5.

In summary, there are several advantages of the proposed framework. First, by segmenting customers, a business can develop specific predictive modeling approaches and marketing strategies to appreciate the value of precise marketing campaigns and increase ROI in marketing. Second, in B2B market, it is essential to identify multi-focal groups of prospects. The predictive modeling task in B2B market is a rare-class learning problem due to low conversion rates. The segmentation process can help to improve the imbalanced class distributions in some segments with higher marketing priorities, and thus can help to produce high quality leads in these segments. Finally, while some segments will have insufficient representative samples and the class distributions of these segments can be more imbalanced, the proposed MTL strategy can leverage the converted cases from some other segments to enhance the overall performances of all the tasks.

4.4.3 Parameter Tuning Process

In the regularization term, there are two hyperparameters, namely, the trade-off parameter λ and the weights ϕ_p ($1 \leq p \leq P$) of the tree structure. To find the optimal values for the hyperparameters, we use the traditional strategy–*grid search*, which is an exhaustive searching through a pre-specified subset of the hyperparameter space. And the grid search approach is guided by the performance metrics (details can be find in 4.5.1), which are evaluated by cross-validation on the training data set. In specific, we exhaustively evaluate all the possible combinations of parameter values sampled from $0, 0.25, \ldots, 0.75, 1$ and $1, 2, \ldots, 9, 10$ with five-fold cross-validation and the best combination is retained.

4.5 Empirical Analysis and Results

4.5.1 Experimental Setup

In this section, we present the empirical study of the multi-focal lead scoring approach. Specifically, we investigate the proposed method using real-world B2B marketing data which capture the probability of prospects to be converted when they interact with marketing campaign activities. As discussed before, the lead scoring model is built using 36 campaign related behavior variables (summarized in Table 4.2) as explanatory variables. The dependent variable is whether the prospect converted or not.

Baselines We compare the performances of five approaches:

- Traditional: one general non-segment approach for all the prospects.

- Single Task Learning (STL): separately built predictive model for each segment.
- MTL-Mean (MTL): Multi-task learning using Mean-induced regularization [23].
- MTL-Level: Multi-task learning using levelwise tree-induced regularization.
- MTL-Node: Multi-task learning using nodewise tree-induced regularization.

The last two approaches are proposed in this work. All the experiments are performed using MATLAB and the MALSAR package for multi-task learning [79]. In the experiments, we perform ten-fold cross validation. Moreover, normalization using *z-score* has been done before training the model.

Performance Metrics Three classic evaluation metrics, namely, precision, recall, F-measure [61], are used for validation. Since we are more interested in the converted class rather than the not-converted class, all these three measures are focus on the prediction of the converted class. The formulas of these three measures are as follows:

$$
\begin{aligned}
Recall &= \frac{\text{number of correct positive predictions}}{\text{total number of positive examples}} \\
Precision &= \frac{\text{number of correct positive predictions}}{\text{total number of positive predictions}} \\
F\text{-}measure &= 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}
\end{aligned}
$$

4.5.2 Performance Comparison

As shown in Figure 4.3, different segments have different conversion rates. Therefore, it is natural for us to use lift chart by the cumulative of ranked top K samples to show the performance of three approaches on each decile, where K equals to the number of conversions in specific segment. Hence, the performance of all four approaches are presented in Figure 4.7. More specifically, the first five rows in the figure represent the results of five segments, respectively. The last row demonstrates the weighted average of performances in five segments, and the weight w is defined as $w = \frac{\text{number of conversions in the segment}}{\text{total number of conversions}}$.

From the weighted average results, we can clearly observe that MTL approaches are superior than traditional non-segment approach and STL approach. In particular, although MTL approaches have similar performances, MTL-Node approach with $\lambda = 0.25$, $\phi_1 = 1$, $\phi_2 = 7$, and $\phi_3 = 4$ yields the consistently better performances than MTL-Mean on all three metrics. In general, both MTL-Level and MTL-Node have much more advantages at the top 30% of the leads across five segments, but MTL-Node consistently reaches better performances at the 100% of total conversion.

As expected, both tree-induced MTL approaches (MTL-Level and MTL-Node) outperform other three approaches in the overall performance. This proves our hypothesis that tree-induced MTL can be used to fine tune the predictive model. STL is not always better than traditional non-segment approach, however the creative integration of MTL in the multi-focal framework can lead to consistent better performances. Also, Segment 3 has the lowest conversion rate of 0.5%. In other words, there are 73 conversions out of 13,456 prospects. Certainly, Segment 3 has insufficient conversion examples to learn the model. However, it is intuitive that similar conversion patterns are shared among different segments, MTL can take advantage of the conversion patterns from other tasks to make up the issue of inadequate representative training samples. In contrast, traditional non-segment approach tends to fail when the number of training examples are small. Therefore, in Segment 3, No-Segment approach yields the worst performance all the time, but MTL-Node can get the final F-measure four times better than that of No-Segment approach.

4.6 Managerial Implications.

4.6.1 Segment Relationships

As we mentioned in Section 4.4.2, ϕ_p represents the weight of p-th parent node, which can interpreted as the behavior similarities among the connected nodes/segments. We also assume that the deeper the tree grows, the larger the value of ϕ_p. This can be validated by the learned optimal values for $\phi_1 = 1 < \phi_3 = 4 < \phi_2 = 7$. The smaller value of ϕ_1 indicates that there is limited commonalities among the five segments, while larger values of ϕ_2 and ϕ_3 show that the commonalities are stronger among the connected segments, especially for the segments connected in ϕ_2. One possible implication for $\phi_2 > \phi_3$ is that segments grouped based on *CompanySize* are more closely related to each other than those grouped based on *Industry*. Thus, it is beneficial to take into account the inner relatedness among some segments when developing new marketing strategies, by using the segment commonality as a general pattern for guidance.

4.6.2 Diversified Campaign Preferences

In Table 4.3, we list the learned coefficients of 36 variables generated by No-Segment, STL and MTL-Node methods in 5 Segments. The positive coefficient represents the

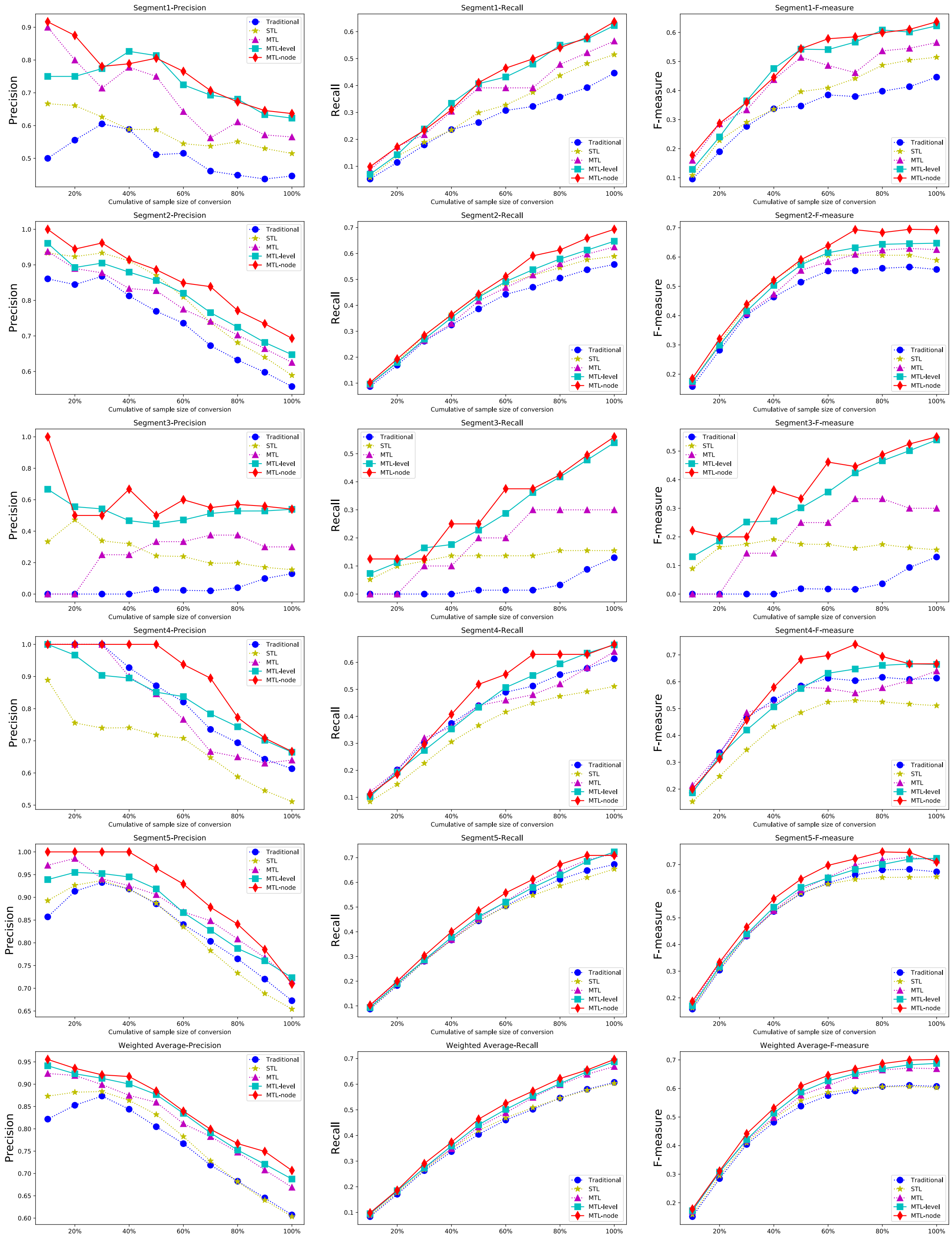


Figure 4.7: The Performance Comparison of Traditional, STL, MTL-Mean, MTL-Level, and MTL-Node.

Table 4.3: Model Coefficient Comparison

Variable	NOSEG	Segment1 STL	Segment1 MTL-Node	Segment2 STL	Segment2 MTL-Node	Segment3 STL	Segment3 MTL-Node	Segment4 STL	Segment4 MTL-Node	Segment5 STL	Segment5 MTL-Node
DaysSinceMindate	-0.9335	-1.1117	-1.1201	-0.6604	-1.1202	-0.4341	-1.1196	-1.6212	-1.5367	-1.6615	-1.5374
Evnt1seminar	0.0989	0.0690	0.1077	0.0953	0.1076	0.2940	0.1080	0.0874	0.1747	0.0695	0.1737
Evnt1corpevent	-0.0315	-0.0911^{ns}	-0.0613	-0.0794	-0.0612	-0.0518	-0.0613	-0.0278	-0.0871	-0.0239	-0.0865
Evnt1webinar	0.0592	0.0836^{ns}	0.0430	0.0152^{ns}	0.0423	-0.0087	0.0429	0.0693^{ns}	0.2174	0.1761	0.2166
$Evnt1conference^{ns}$	0.0041	-0.0269	-0.0085	0.0037	-0.0084	-0.037	-0.0085	0.0014	-0.0119	-0.0215	-0.0121
Evnt1tradeshow	-0.1463	-0.2643^{ns}	-0.2139	-0.2524^{ns}	-0.2137	-0.1107^{ns}	-0.2138	-0.1390^{ns}	-0.0654^{ns}	-0.1266^{ns}	-0.0659
Evnt1attended	0.0194^{ns}	0.0235^{ns}	0.0564	0.1178^{ns}	0.0574	0.1235^{ns}	0.0560	0.0263^{ns}	-0.0028	0.0273^{ns}	-0.0032
$Evntnumconferences^{ns}$	0.0022	-0.0292	-0.0139	0.0113	0.0139	-0.0366	-0.0139	-0.0120	-0.0126	-0.0231	-0.0128
$Evntnumcorpevents^{ns}$	0.0315	-0.0369	0.0236	0.0183	0.0237	-0.0405	0.0235	0.0313	0.0709	0.0289	0.0716
Evntnumseminars	0.0317	0.0104^{ns}	0.0212^{ns}	0.0429^{ns}	0.0215^{ns}	0.1261^{ns}	0.0214^{ns}	0.0284^{ns}	0.0638	0.0144^{ns}	0.0631
Evntnumtradeshows	-0.1399^{ns}	-0.2551^{ns}	-0.2008	-0.1517^{ns}	-0.2005	-0.1083^{ns}	-0.2006	-0.1388^{ns}	-0.0630	-0.0466^{ns}	-0.0631
Evntnumwebinars	-0.0316	-0.0323^{ns}	-0.0563^{ns}	-0.0266^{ns}	-0.0574^{ns}	-0.0259^{ns}	-0.0573^{ns}	-0.0309^{ns}	-0.0380	-0.0280^{ns}	-0.0374
Evntnumtotal	0.0003	-0.0583^{ns}	-0.0214^{ns}	-0.0321^{ns}	-0.0213^{ns}	-0.0128^{ns}	-0.0214^{ns}	0.0024^{ns}	0.0336^{ns}	0.0047^{ns}	0.0335^{ns}
Offer1OfficialWebsite	0.1191	0.2204^{ns}	0.1884	0.1970^{ns}	0.1881	0.2323^{ns}	0.1883	0.1303^{ns}	0.1892	0.1529	0.1884
Offer1email	-0.0132^{ns}	-0.0920^{ns}	-0.0719	-0.0761	-0.0720	-0.0012^{ns}	-0.0718	0.0474^{ns}	0.0597^{ns}	0.0504	0.0590^{ns}
Offer1sem	-0.0126	-0.0011^{ns}	-0.0264^{ns}	-0.0246^{ns}	-0.0263^{ns}	-0.0510^{ns}	-0.0265^{ns}	-0.0411^{ns}	-0.0641	-0.1798^{ns}	-0.0643
Offer1webads	-0.3526	-0.4427^{ns}	-0.4429	-0.3258^{ns}	-0.4424	-0.2837^{ns}	-0.4427	-0.4114^{ns}	-0.3470	-0.4124^{ns}	-0.3496
Offer1directmail	-0.0374^{ns}	-0.0652^{ns}	-0.0565	-0.0516	-0.0564	-0.0311^{ns}	-0.0564	-0.0473^{ns}	-0.0744	-0.0758	-0.0743
Offer1take	-0.3168	-0.2799	-0.3724	-0.1824^{ns}	-0.3723	-0.1762^{ns}	-0.3727	-0.3647^{ns}	-0.4080	-0.4095	-0.4079
OffernumOfficialWebsite	0.0687	0.0739	0.0709	0.0733	0.0708	0.1849^{ns}	0.0709	0.1188^{ns}	0.1468	0.1433	0.1463
Offernumdirectmail	-0.0380^{ns}	-0.0635	-0.0556	-0.0531^{ns}	-0.0563	-0.0304^{ns}	-0.0553	-0.0822^{ns}	-0.0775	-0.0841^{ns}	-0.0774
Offernumemail	0.0815^{ns}	0.1111	0.0612	0.0427^{ns}	0.0609	-0.0164^{ns}	0.0610	0.0966^{ns}	0.2222	0.1181^{ns}	0.2213
$Offernumsem^{ns}$	0.0136	-0.0211	0.0273	0.0052	0.0287	-0.0459	0.0285	-0.1823	0.0032	-0.1826	0.0027
Offernumwebads	-0.2664	-0.3618	-0.2974	-0.2859^{ns}	-0.2968	-0.2397^{ns}	-0.2973	-0.2242^{ns}	-0.2489	-0.1539	-0.2483
Offernumtotal	-0.0818^{ns}	-0.0665	-0.0849	-0.1495^{ns}	-0.0849	-0.1608^{ns}	-0.0850	-0.0375^{ns}	0.0002^{ns}	0.0017^{ns}	-0.0001^{ns}
Prod1trial	-0.2104	-0.1645	-0.1137	-0.1819	-0.1135	-0.1250	-0.1137	-0.2586	-0.3747	-0.3662	-0.3741
Prod1trialfree	0.2813	0.4076	0.3478	0.1457	0.3475	0.1183	0.3477	0.2638	0.2913	0.2331	0.2908
Prodnumtotal	-0.182	-0.1987	-0.1420	-0.0647	-0.1419	-0.056	-0.1419	-0.2207	-0.3133	-0.2268	-0.3122
Prodnumtotalfree	0.0632	0.0915	0.0681	0.0345	0.0680	0.0011	0.0681	0.0472	0.0333	-0.0082	0.0335
Activity1unsubscribe	-0.1132	-0.0832^{ns}	-0.0662	-0.0230	-0.0662	-0.0162^{ns}	-0.0659	-0.1500	-0.1424	-0.1804	-0.1439
Activitynumunsubscribetotal	-0.1032^{ns}	-0.0857^{ns}	-0.0571	-0.0366^{ns}	-0.0571	0.0202^{ns}	-0.0568	-0.1812^{ns}	-0.1630	-0.1998^{ns}	-0.1623
Actiitynumtotal	-0.0189	-0.1162	-0.0442	-0.0416	-0.0441	-0.0407	-0.0441	0.1017	0.1165	0.1396	0.1157
NumInt6Month	-0.0915^{ns}	-0.0748	-0.0704	-0.0801	-0.0704	-0.0537	-0.0704	-0.1226	-0.2131	-0.1379	-0.2113
NumCnt6Month	-0.0957	-0.0749	-0.0759	-0.0738	-0.0759	-0.0535^{ns}	-0.0759	-0.1308	-0.2057	-0.1899	-0.2044
NumInt3Month	-0.0134^{ns}	-0.0592^{ns}	-0.0452	-0.0182	-0.0451	-0.0513^{ns}	-0.0452	0.0632^{ns}	-0.0262^{ns}	0.1109	-0.0244^{ns}
NumCnt3Month	-0.0088	-0.067^{ns}	-0.0591	-0.0353^{ns}	-0.0591	-0.0516^{ns}	-0.0591	0.1024^{ns}	0.0528^{ns}	0.1896	0.0546^{ns}

*All estimates are statistically significant at 0.001 level, except the ones marked with *ns*.

positive relationship between the variable and the expected result. A comparison of the signs of the coefficients tells why there is a difference in the performances of these three approaches. For example, there are several opposite signs of coefficients between the results of STL and that of MTL-Tree, such as the coefficients of *Evntnumcorpevents*(Corporate events) and *Offernumsem* (Offer from search engine marketing) are negative in STL models but are positive in MTL-Node. Positive relationships showed in MTL models seem more reasonable because corporate events and offers from search engine marketing are quite influential and efficient in practice. On the other hand, the coefficients of *Offernumtotal* are negative in MTL models but are positive in STL models. It is commonly understood that a large number of clicked offers from all kinds of sources does not necessarily lead to the high volume of conversions, and thus there is no positive relationship between these two incidents.

Furthermore, we discuss the shared effective campaign types across segments with different conversion rates. To study the effectiveness of a specific campaign, the absolute value of the coefficient of the campaign variable reflects the influence of that variable to some extent. For example, *Prod1trialfree* ("free product trial") has the highest positive value across segments. This indicates that "free product trial" is the most significant campaign to generate more conversion cases. Also, the followed positive influential campaign types are *Evnt1seminar* (seminar), *Offernumemail* (offer from email), *OffernumOfficialwebsite* (offer from company official website), and *Evnt1webinar* (webinar). This may suggest that for the the event campain type, Seminar is relatively more effective than Webinar since the former contains more . While for the offer campaign type, the offers from email and official website are more effective than that from search engine ads and other website ads, possibly due to the perceived higher quality communication via its corporate advertising. Since *Web ads* related offer campaign variables have relatively high negative coefficients, we can infer that *Web ads* related offer campaigns have insignificant or even unfavorable influence on boosting the conversion rate. In short, from the analysis of the MTL-Node model we can find the pattern of effective campaigns shared among different segments. This understanding can help the marketing managers to optimize investment on more efficacious campaigns.

In addition, we also show the significance for all the model coefficient estimates. All estimates are statistically significant at 0.01 level, except for those denoted as *ns*. Specifically, *Evnt1conference*, *Evntnumconference*, and *Offernumsem* appear to have relatively less significance across all the segments. This is consistent with the fact that conference is a rare campaign type with relatively fewer attendants. Moreover, a key insight is that there are different preferences between new customers (Segment 1-3) and existing customers (Segment 4-5) due to the different significant levels in *Offer1Email* and *Evntnumcorpevents*. The email campaign type is more important in the prediction models for new customers than existing customers. In contrast, corporate events are more critical for the existing customers than new customers. Also, in the STL, variables related to the count of attended offer campaign types (such as *Offernumemail*, *Offernumsem*) are meaningful addition to the prediction model for customers in Segment 1 but not in the model for Segment 2. These variables however are significant in our MTL-Node models.

4.7 Summary

In this work, we developed a multi-focal leading scoring approach to assist B2B marketing from two aspects. First, it can help to prioritize a large collection of prospects into a high-quality lead list. Second, it can help to develop customized segment-specific marketing campaign strategies. Indeed, conventional lead scoring models are largely rely on marketers' experiences and business rules, while there are significant marketing insights buried in the large collection of B2B marketing data. Thus, it is critical for us to understand how data-driven lead scoring techniques can help to increase conversion rates, general sales pipeline, and improve marketing ROI.

Specifically, in the multi-focal learning framework, leads are first divided into several different focal groups (segments) based on the CHAID decision tree. Then, a lead scoring model is built for each segment by multi-task learning. There are both technical and marketing advantages of our approach. In terms of technical advantages, the multi-focal learning can directly improve the learning performances of the segments which have a more balanced class distribution due to segmentation. Also, the creative integration of multi-task learning into the multi-focal learning framework can help to improve the learning performances in segments which are short of representative samples due to segmentation. In terms of marketing advantages, the multi-focal learning allows us to directly increase conversion rates in high-quality customer segments. Also, with the help of MTL, it helps to improve the overall marketing ROI as well.

As a real-world deployment, we applied the multi-focal lead scoring model for B2B marketing in a large database containing a large amount of prospects' demographic and behavior information. The experimental studies show that our approach leads to better learning performances than conventional lead scoring methods.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this chapter, we conclude the dissertation first with an overall review and then a general discussion about future research directions.

5.1 Review of Major Results

This thesis addresses real-world B2B marketing challenges and develop several data-intensive application for marketing anlaytics by effectively modeling with various customer data. More specifically, the research work has yielded the following results.

B2B marketing campaign recommendation. In this work, we developed a novel recommender system to combine the temporal and social factors captured in customer campaign participating records and the customer community networks for B2B marketing campaign recommendations. The goal is to provide the marketers with a better marketing strategy to expedite the customer conversion cycle and boost the customer conversion ratio. The proposed B2B marketing campaign recommender system strategically integrates the temporal preferences, community relationships and the campaign interestingness in the framework. Specifically, we first represented the information-rich temporal content in customer behavior records using the temporal graphs. Next, with the personalized temporal graphs, we computed the low-rank graph reconstruction to predict the unobserved graph edges. Moreover, we regularized the graph reconstruction with the community network of the business customers. Furthermore, the proposed approach extended the preliminary work by considering the skewed campaign frequency data characteristic of B2B marketing data, to incorporate the campaign interestingness information in the model. Finally, we developed efficient algorithms to compute the optimal solutions, which we have applied on several real-world B2B marketing data sets.

Buyer targeting optimization. Due to the heterogeneity across the customer groups, customer segmentation and buyer targeting these two essential marketing tasks are combined in a simple step-by-step way for tailored marketing strategies. It is still unclear how to implement these two tasks in a more integrated and optimized way. To this end, we first provided a novel mathematical formulation to integrate customer segmentation and buyer targeting into a unified optimization problem, so that the customers are grouped into segments where the promising buyers can be most easily identified. In other words, these two tasks are performed simultaneously in a unified optimization framework which combines the clustering and classification objectives. To solve the optimization problem, we developed an iterative K-Classifiers Segmentation algorithm, where the customer segments are formed with customers' buyer targeting models. Moreover, we showed that the segmentation results can also be consistent with the features on customer profiles. Finally, we applied our approach on both synthetic data and real-world Business-to- Business (B2B) marketing scenarios. In addition to targeting (classification) accuracies, experiment results showed that our approach can provide interpretable customer segmentation solutions and reveals new marketing insights.

Multi-focal lead scoring. In this work, we developed a multi-focal leading scoring approach to help not only prioritize a large collection of prospects into a high-quality lead list, but also develop customized segment-specific marketing campaign strategies. Specifically, leads are first divided into several different focal groups (segments) based on the CHAID decision tree. Then, a lead scoring model is built for each segment with multi-task learning. The inventive integration of multi-task learning into the multi-focal learning framework can help to improve the learning performances in segments which are short of representative samples due to segmentation. Finally, we applied the multi-focal lead scoring model on a real-world B2B marketing dataset. The experimental studies show that our approach gives a better insight into factors influencing the conversion of leads.

5.2 Future Work

The big data accumulated in marketing industry has greatly changed the paradigm of marketing services. Data-driven solutions are needed for effective knowledge discovery from large scale marketing datasets. In general, my current and future research focus will still be on the fundamental analytical challenges of big marketing data and the discovery of in-depth business insights. I would like to contribute to the systematic research on business intelligence, particularly from the perspectives of marketing analytics, and personalized recommendation, to combine business data with time series, mobile, and personalized dimensions, to make informed decisions, to reduce managerial risk, and finally, to boost return on investment. Towards this direction, I will focus on the following three topics: (i) B2B buyer stage and cycle time analysis, (ii) customer churn and retention analysis, and (iii) product cross-selling/up-selling recommender system. For instance, it is of significant business value if we can discover characteristic and critical purchase stages from B2B customer behavior observations, and if we can visually display the important stages of the buying process. To this end, we can apply data mining techniques such as clustering and association analysis to derive the temporal relationships among the customer's behavior and discover the significant indicators of stage transition during the process.

References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *TKDE*, *17*(6), 734–749.

Antipov, E., & Pokryshevskaya, E. (2010). Applying chaid for logistic regression diagnostics and classification accuracy improvement. *Journal of Targeting, Measurement and Analysis for Marketing*, *18*(2), 109–117.

Apte, C., Bibelnieks, E., Natarajan, R., Pednault, E., Tipu, F., Campbell, D., & Nelson, B. (2001). Segmentation-based modeling for advanced targeted marketing. In *Proceedings of the seventh acm sigkdd international conference on knowledge discovery and data mining* (pp. 408–413).

Bakker, B., & Heskes, T. (2003). Task clustering and gating for bayesian multitask learning. *The Journal of Machine Learning Research*, *4*, 83–99.

Batra, R., & Shepard, D. (1995). *The new direct marketing: How to implement a profit-driven database marketing strategy*. Irwin Professional.

Bertsimas, D., Mersereau, A. J., & Patel, N. R. (2003). Dynamic classification of online customers. In *Sdm* (pp. 107–118).

Boone, D. S., & Roehm, M. (2002). Retail segmentation using artificial neural networks. *International journal of research in marketing*, *19*(3), 287–301.

Bose, I., & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, *195*(1), 1–16.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Uai* (pp. 43–52).

Bubeck, S., & Luxburg, U. v. (2009). Nearest neighbor clustering: A baseline method for consistent clustering with arbitrary objective functions. *The Journal of Machine Learning Research*, *10*, 657–698.

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science*, *14*(4), 378–394.

Campos, P. G., Díez, F., & Cantador, I. (2014). Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *UMUAI*, *24*(1-2), 67–119.

Canny, J. (2004). Gap: a factor model for discrete data. In *Sigir* (pp. 122–129).

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41–75.

Chen, W., Hsu, W., & Lee, M. L. (2013). Modeling user's receptiveness over time for recommendation. In *Sigir* (pp. 373–382).

Cheng, B., & Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical science*, 2–30.

Cheng, C., Yang, H., Lyu, M. R., & King, I. (2013). Where you like to go next: Successive point-of-interest recommendation. In *Ijcai.*

Chou, P. B., Grossman, E., Gunopulos, D., & Kamesam, P. (2000). Identifying prospective customers. In *Proceedings of the sixth acm sigkdd international conference on knowledge discovery and data mining* (pp. 447–456).

Chua, F. C. T., Lauw, H. W., & Lim, E.-P. (2011). Predicting item adoption using social correlation. In *Sdm* (Vol. 11).

Currim, I. S. (1981). Using segmentation approaches for better prediction and understanding from consumer mode choice models. *Journal of Marketing Research*, 301–309.

Dolnicar, S. (2003). Using cluster analysis for market segmentation-typical misconceptions, established methodological weaknesses and some recommendations for improvement.

etal., N. L. (2010). Temporal diversity in recommender systems. In *Sigir* (pp. 210–217).

Evgeniou, T., & Pontil, M. (2004). Regularized multi–task learning. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 109–117).

Fournier-Viger, P., Wu, C.-W., Tseng, V. S., Cao, L., & Nkambou, R. (2015). Mining partially-ordered sequential rules common to multiple sequences. *TKDE*, *27*(8), 2203–2216.

Ge, Y., Xiong, H., Zhou, W., Li, S., & Sahoo, R. (2011). Multifocal learning for customer problem analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 24.

Grbovic, M., Djuric, N., & Vucetic, S. (2012). Supervised clustering of label ranking data. In *Sdm* (pp. 94–105).

Gupta, S., & Chintagunta, P. K. (1994). On using demographic variables to determine segment membership in logit mixture models. *Journal of Marketing Research*, 128–136.

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., … Sriram, S. (2006). Modeling customer lifetime value. *Journal of Service Research*, *9*(2), 139–155.

Haughton, D., & Oulabi, S. (1997). Direct marketing modeling with cart and chaid. *Journal of Interactive Marketing*, *11*(4), 42–52.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *JMLR*, *5*, 1457–1469.

Jonker, J.-J., Piersma, N., & Potharst, R. (2006). A decision support system for direct mailing decisions. *Decision support systems*, *42*(2), 915–925.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, 119–127.

Kestnbaum, R. D., Kestnbaum, K. T., & Ames, P. W. (1998). Building a longitudinal contact strategy. *Journal of Interactive Marketing*, *12*(1), 56–62.

Kim, Y., Street, W. N., Russell, G. J., & Menczer, F. (2005). Customer targeting: A neural network approach guided by genetic algorithms. *Management Science*, *51*(2), 264–276.

Koren, Y. (2010). Collaborative filtering with temporal dynamics. *Communications of the ACM*, *53*(4), 89–97.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*(8), 30–37.

Krieger, A. M., & Green, P. E. (1996). Modifying cluster-based segments to enhance agreement with an exogenous response variable. *Journal of Marketing Research*, 351–363.

Levin, N., & Zahavi, J. (2001). Predictive modeling using segmentation. *Journal of Interactive Marketing*, *15*(2), 2–22.

Liu, B., Kong, D., Cen, L., Gong, N. Z., Jin, H., & Xiong, H. (2015). Personalized mobile app recommendation: Reconciling app functionality and user privacy preference. In *Wsdm* (pp. 315–324).

Liu, B., Xiong, H., Papadimitriou, S., Fu, Y., & Yao, Z. (2015). A general geographical probabilistic factor model for point of interest recommendation. *TKDE*, *27*(5), 1167–1179.

Liu, C., Hu, T., Ge, Y., & Xiong, H. (2012). Which distance metric is right: An evolutionary k-means view. In *Sdm* (pp. 907–918).

Liu, C., Zhang, K., Xiong, H., Jiang, G., & Yang, Q. (2014). Temporal skeletonization on sequential data: Patterns, categorization, and visualization. In *Sigkdd.*

Liu, N. N., He, L., & Zhao, M. (2013). Social temporal collaborative ranking for context aware movie recommendation. *TIST*, *4*(1), 15.

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., & Wu, S. (2013). Understanding and enhancement of internal clustering validation measures. *IEEE transactions on cybernetics*, *43*(3), 982–994.

Lu, Y., Tsaparas, P., Ntoulas, A., & Polanyi, L. (2010). Exploiting social context for review quality prediction. In *Www.*

Ma, H., King, I., & Lyu, M. R. (2009). Learning to recommend with social trust ensemble. In *Sigir* (pp. 203–210).

Ma, H., Lyu, M. R., & King, I. (2009). Learning to recommend with trust and distrust relationships. In *Recsys* (pp. 189–196).

Ma, H., Zhou, D., Liu, C., Lyu, M. R., & King, I. (2011). Recommender systems with social regularization. In *Wsdm* (pp. 287–296).

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of rfm, chaid, and logistic regression. *Journal of business research*, *60*(6), 656–662.

Mnih, A., & Salakhutdinov, R. (2007). Probabilistic matrix factorization. In *Nips* (pp. 1257–1264).

Morwitz, V. G., & Schmittlein, D. (1992). Using segmentation to improve sales forecasts based on purchase intent: Which" intenders" actually buy? *Journal of Marketing Research*, 391–405.

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

Pálovics, R., Benczúr, A. A., Kocsis, L., Kiss, T., & Frigó, E. (2014). Exploiting temporal influence in online recommendation. In *Recsys* (pp. 273–280).

Pang-Ning, T., Steinbach, M., Kumar, V., et al. (2006). Introduction to data mining. In *Library of congress*.

Paterek, A. (2007). Improving regularized singular value decomposition for collaborative filtering. In *Kdd cup and workshop* (Vol. 2007, pp. 5–8).

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: review and suggestions for application. *Journal of marketing research*, 134–148.

Rendle, S., Freudenthaler, C., & Schmidt-Thieme, L. (2010). Factorizing personalized markov chains for next-basket recommendation. In *Www* (pp. 811–820).

Reutterer, T., Mild, A., Natter, M., & Taudes, A. (2006). A dynamic segmentation approach for targeting and customizing direct marketing campaigns. *Journal of Interactive Marketing*, *20*(3), 43–57.

Saunders, J. (1980). Cluster analysis for market segmentation. *European Journal of Marketing*, *14*(7), 422–435.

Shen, Y., & Jin, R. (2012). Learning personal+ social latent factor model for social recommendation. In *Sigkdd* (pp. 1303–1311).

Tan, P.-N., et al. (2006). *Introduction to data mining.* Pearson Education India.

Tang, X., & Zhou, J. (2013). Dynamic personalized recommendation on sparse data. *TKDE*, *25*(12), 2895–2899.

Turpin, A., & Scholer, F. (2006). User performance versus precision measures for simple search tasks. In *Sigir.*

Vargo, S. L., & Lusch, R. F. (2004). Evolving to a new dominant logic for marketing. *Journal of marketing*, 1–17.

Vellido, A., Lisboa, P., & Meehan, K. (1999). Segmentation of the on-line shopping market using neural networks. *Expert Systems with Applications*, *17*(4), 303–314.

Verhoef, P. C., Spring, P. N., Hoekstra, J. C., & Leeflang, P. S. (2003). The commercial use of segmentation and predictive modeling techniques for database marketing in the netherlands. *Decision Support Systems*, *34*(4), 471–481.

Wang, J., & Zhang, Y. (2013). Opportunity model for e-commerce recommendation: right product; right time. In *Sigir* (pp. 303–312).

Wilson, R. D. (2003). Using online databases for developing prioritized sales leads. *Journal of Business & Industrial Marketing*, *18*(4/5), 388–402.

Wu, J., Xiong, H., Wu, P., & Chen, J. (2007). Local decomposition for rare class analysis. In *Proceedings of the 13th acm sigkdd international conference on knowledge discovery and data mining* (pp. 814–823).

Xiang, L., Yuan, Q., Zhao, S., Chen, L., Zhang, X., Yang, Q., & Sun, J. (2010). Temporal recommendation on graphs via long-and short-term preference fusion. In *Sigkdd.*

Xiong, L., Chen, X., Huang, T.-K., Schneider, J. G., & Carbonell, J. G. (2010). Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *Sdm* (Vol. 10, pp. 211–222).

Yang, J., Liu, C., Teng, M., & Xiong, H. (2015). Exploiting temporal and social factors for b2b marketing campaign recommendations. In *Icdm.*

Yang, X., Steck, H., & Liu, Y. (2012). Circle-based recommendation in online social networks. In *Sigkdd* (pp. 1267–1275).

Yap, G.-E., Li, X.-L., & Philip, S. Y. (2012). Effective next-items recommendation via personalized sequential pattern mining. In *Dasfaa.*

Yin, H., Cui, B., Chen, L., Hu, Z., & Zhou, X. (2015). Dynamic user modeling in social media systems. *TOIS*.

Zhang, J.-D., Chow, C.-Y., & Li, Y. (2014). Lore: Exploiting sequential influence for location recommendations. In *Sigspatial.*

Zhao, G., Lee, M. L., Hsu, W., & Chen, W. (2012). Increasing temporal diversity with purchase intervals. In *Sigir* (pp. 165–174).

Zhao, G., Lee, M. L., & Wynne, H. (2014). Utilizing purchase intervals in latent clusters for product recommendation. In *Sna-kdd* (p. 4).

Zhou, J., Chen, J., & Ye, J. (2012). Malsar: Multi-task learning via structural regularization. *Arizona State Univ*.

APPENDIX

.1 Aggregation Rules of Demographic Variables

Table 1: Aggregation Rules of Demographic Variables

Term	Definition
New	Contact from an account with the account status of not “Current” or “Former”
Existing	Contact from an account with the account status of “Current”
Enterprise	Account with more than 500 employees
SMB	Account with less than or equal to 500 employees
IT Staff	{Analyst, Network Administrator/Engineer, Architect, System Administrator}
IT Management	{IT Manager, Director}
Executive	{C-Level / President / VP}
Researcher	{Student, Owner / Partner, Consultant}
Other	{Assistant, Developer / Programmer, Business Manager, Other}
Heavy Hitter	High volume and high conversion industry: Healthcare, Finance, Retail, Education, Manufacturing, Government
Laggard	High volume and low conversion industry : Business Services, Technology, Business Consulting
Potential	Low volume and high conversion industry

.2 Mathematical Derivations

We use gradient descent procedures to iteratively update the optimization variables: W, c, λ, and ϕ_p for $p = 1, \cdots, P$. The element-wise gradients are given below, but note that the gradient updating can be vectorized for improved efficiency (e.g., in MATLAB/Numpy).

$$
\begin{aligned}
\frac{\partial \Omega(w)}{\partial \lambda} &= \frac{\partial \lambda \sum_{i=1}^{T} \sum_{j=1}^{T} G_{ij} ||W^{(i)} - W^{(j)}||^2}{\partial \lambda} \\
&= \sum_{i=1}^{T} \sum_{j=1}^{T} G_{ij} ||W^{(i)} - W^{(j)}||^2
\end{aligned}
\tag{1}
$$

$$
\begin{aligned}
\frac{\partial \Omega(w)}{\partial \phi_p} &= \frac{\partial \lambda \sum_{i=1}^{T} \sum_{j=1}^{T} G_{ij} ||W^{(i)} - W^{(j)}||^2}{\partial \phi_p} \\
&= \frac{\partial \lambda \sum_{i=1}^{T} \sum_{j=1}^{T} \sum_{l=1}^{L} \phi_1 G_l ||W^{(i)} - W^{(j)}||^2}{\partial \phi_p} \\
&= \frac{\lambda \sum_{i=1}^{T} \sum_{j=1}^{T} ||W^{(i)} - W^{(j)}||^2 \partial (\sum_{l=1}^{L} G_l \phi_p)}{\partial \phi_p} \\
&= \lambda \sum_{i=1}^{T} \sum_{j=1}^{T} G_p ||W^{(i)} - W^{(j)}||^2
\end{aligned}
\tag{2}
$$

$$
\begin{aligned}
\frac{\partial L(w,c)}{\partial c} &= \frac{\partial \sum_{t=1}^{T} \sum_{n=1}^{N_t} log(1 + e^{-y_n^{(t)}(x_n^{(t)} w^{(t)} + c^{(t)})})}{\partial c} \\
&= \frac{\sum_{t=1}^{T} \sum_{n=1}^{N_t}}{1 + e^{-y_n^{(t)}(x_n^{(t)} w^{(t)} + c^{(t)})}} \frac{\partial (1 + e^{-y_n^{(t)}(x_n^{(t)} w^{(t)} + c^{(t)})})}{\partial c} \\
&= \frac{\sum_{t=1}^{T} \sum_{n=1}^{N_t}}{1 + e^{-y_n^{(t)}(x_n^{(t)} w^{(t)} + c^{(t)})}} \frac{\partial (e^{-y_n^{(t)}(x_n^{(t)} w^{(t)} + c^{(t)})})}{\partial c} \\
&= \frac{\sum_{t=1}^{T} \sum_{n=1}^{N_t} e^{-y_n^{(t)}(x_n^{(t)} w^{(t)} + c^{(t)})}}{1 + e^{-y_n^{(t)}(x_n^{(t)} w^{(t)} + c^{(t)})}} \frac{\partial (-y_n^{(t)}(x_n^{(t)} w^{(t)} + c^{(t)}))}{\partial c} \\
&= \frac{\sum_{t=1}^{T} \sum_{n=1}^{N_t} e^{-y_n^{(t)}(x_n^{(t)} w^{(t)} + c^{(t)})} * e^{-y_n^{(t)}}}{1 + e^{-y_n^{(t)}(x_n^{(t)} w^{(t)} + c^{(t)})}}
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
\frac{\partial L(w,c)}{\partial w} &= \frac{\partial \sum_{t=1}^{T}\sum_{n=1}^{N_t} log(1+e^{-y_n^{(t)}(x_n^{(t)}w^{(t)}+c^{(t)})})}{\partial w} \\
&= \frac{\sum_{t=1}^{T}\sum_{n=1}^{N_t}}{1+e^{-y_n^{(t)}(x_n^{(t)}w^{(t)}+c^{(t)})}}\frac{\partial(1+e^{-y_n^{(t)}(x_n^{(t)}w^{(t)}+c^{(t)})})}{\partial w} \\
&= \frac{\sum_{t=1}^{T}\sum_{n=1}^{N_t}}{1+e^{-y_n^{(t)}(x_n^{(t)}w^{(t)}+c^{(t)})}}\frac{\partial(e^{-y_n^{(t)}(x_n^{(t)}w^{(t)}+c^{(t)})})}{\partial w} \\
&= \frac{\sum_{t=1}^{T}\sum_{n=1}^{N_t} e^{-y_n^{(t)}(x_n^{(t)}w^{(t)}+c^{(t)})}}{1+e^{-y_n^{(t)}(x_n^{(t)}w^{(t)}+c^{(t)})}}\frac{\partial(-y_n^{(t)}(x_n^{(t)}w^{(t)}+c^{(t)}))}{\partial w} \\
&= \frac{\sum_{t=1}^{T}\sum_{n=1}^{N_t} e^{-y_n^{(t)}(x_n^{(t)}w^{(t)}+c^{(t)})} * e^{-y_n^{(t)}x_n^{(t)}}}{1+e^{-y_n^{(t)}(x_n^{(t)}w^{(t)}+c^{(t)})}}
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
\frac{\partial \Omega(w)}{\partial w_{i_0j_0}} &= \frac{\partial\lambda\sum_{i=1}^{T}\sum_{j=1}^{T} G_{ij}||W^{(i)}-W^{(j)}||^2}{\partial w_{i_0j_0}} \\
&= \frac{\lambda\partial\sum_{i=j}^{T} G_{ij}||W^{(i)}-W^{(j)}||^2}{\partial w_{i_0j_0}} \\
&= \frac{\lambda\partial\sum_{j=1}^{T} G_{i_0j}||W^{(i_0)}-W^{(j)}||^2+\lambda\partial\sum_{i=1}^{T} G_{ii_0}||W^{(i_0)}-W^{(j)}||^2}{\partial w_{i_0j_0}} \\
&= \frac{\lambda\partial\sum_{k=1}^{T} G_{i_0j}(W_{i_0k}-W_{jk})^2+\lambda\partial\sum_{k=1}^{T} G_{ii_0}(W_{ik}-W_{i_0k})^2}{\partial w_{i_0j_0}} \\
&= \lambda\sum_{j=1}^{T} G_{i_0j} * 2(W_{i_0j_0}-W_{jj_0}) + \lambda\sum_{i=1}^{T} G_{ii_0} * 2(W_{i_0j_0}-W_{ij_0})
\end{aligned} \tag{5}
$$