ABSTRACT

STATISTICAL MACHINE LEARNING METHODS FOR MINING SPATIAL AND TEMPORAL DATA

by
Fei Tan

Spatial and temporal dependencies are ubiquitous properties of data in numerous domains. The popularity of spatial and temporal data mining has thus grown with the increasing prevalence of massive data. The presence of spatial and temporal attributes not only provides complementary useful perspectives, but also poses new challenges to the representation and integration into the learning procedure. In this dissertation, the involved spatial and temporal dependencies are explored with three genres: sample-wise, feature-wise, and target-wise. A family of novel methodologies is developed accordingly for the dependency representation in respective scenarios.

First, dependencies among discrete, continuous and repeated observations are studied using illustrative examples in urban computing and video clicks. Specifically, discrete Markov random field and time-aware latent hierarchical models are developed to capture the underlying spatiotemporal interactions among different spots. In addition, an item-specific effect aware method is proposed to model consistent effects involved in repeated observational records. Second, feature-wise spatiotemporal interactions are investigated under the framework of deep learning with applications to genomic sequences and audience logs. Regarding spatial dependency among homogeneous features (e.g., genomic sequence), a customized convolutional neural network is leveraged to capture underlying motifs formed by spatial interactions.

To advance the characterization of spatiotemporal interactions among heterogeneous features, a blended learning scheme is established to keep track of the evolution of involved patterns. For both feature-wise dependencies, a saliency maps based context analysis protocol is introduced to interpret and visualize the manner how spatial-temporal attributes are associated with target responses. Lastly, this dissertation covers the temporal dependence of response target variables with applications to competing risks in financial loans. A hierarchical grading framework is proposed to integrate two risks of loans both qualitatively and quantitatively based on temporal constraints. The framework is then divided into multiple binary classification sub-problems. All of the proposed methods are evaluated by systematic experiments based on synthetic data and real-world data repositories in various scenarios. The empirical results demonstrate the appealing performance in different regards.

Taken together, this dissertation elucidates spatiotemporal data from three perspectives and is dedicated to developing desirable and feasible schemes for the representation of spatial and temporal mechanisms.

STATISTICAL MACHINE LEARNING METHODS FOR MINING SPATIAL AND TEMPORAL DATA

by
Fei Tan

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

May 2019

APPROVAL PAGE

STATISTICAL MACHINE LEARNING METHODS FOR MINING SPATIAL AND TEMPORAL DATA

Fei Tan

Zhi Wei, Dissertation Advisor Date
Associate Professor of Computer Science, NJIT

James M. Calvin, Committee Member Date
Professor of Computer Science, NJIT

Senjuti Basu Roy, Committee Member Date
Assistant Professor of Computer Science, NJIT

NhatHai Phan, Committee Member Date
Assistant Professor of Informatics, NJIT

Wenge Guo, Committee Member Date
Associate Professor of Mathematical Sciences, NJIT

BIOGRAPHICAL SKETCH

Author:	Fei Tan
Degree:	Doctor of Philosophy
Date:	May 2019

Undergraduate and Graduate Education:

- Ph.D. in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2019
- M.Sc. in Electronic Engineering,
  Zhejiang University, Hangzhou, China, 2014
- B.Eng. in Electronic Engineering,
  Shandong University of Science and Technology, Tsingtao, China, 2011

Major:	Computer Science

Presentations and Publications:

F. Tan, Z. Wei, A. Pani, and Z. Yan, User Response Driven Content Understanding with Causal Inference, *Under Review*

F. Tan, T. Tian, X. Hou, Z. Wei, L. Gu, and H. Hakonarson, Elucidation of DNA Methylation on N^6-Adenine with Deep Learning, Nature Machine Intelligence, *Under Review*

F. Tan, C. Cheng, and Z. Wei, Modeling and Elucidation of Housing Price, Data Mining and Knowledge Discovery, DOI: 10.1007/s10618-018-00612-0, 2019 (IF$_{2017}$=2.481)

F. Tan, X. Hou, J. Zhang, Z. Wei, and Z. Yan, A Deep Learning Approach to Competing Risks Representation in Peer-to-Peer Lending, IEEE Transactions on Neural Networks and Learning Systems, DOI: 10.1109/TNNLS.2018.2870573, 2018 (IF$_{2017}$=7.982)

F. Tan, X. Hou, J. Zhang, Z. Wei, Z. Yan, and S. Weng, A Novel Risk Assessment Scheme and Practice for Peer-to-Peer Lending, SIGKDD Workshop Data Science in Fintech, London, UK, August 2018

F. Tan, Z. Wei, J. He, X. Wu, B, Peng, H. Liu, and Z. Yan, A Blended Deep Learning Approach for Predicting User Intended Actions, 18^{th} IEEE International Conference on Data Mining, Singapore, November 2018 (Acceptance Rate: 84/948=8.86%)

F. Tan, K. Du, Z. Wei, H. Liu, C. Qin, and R. Zhu, Modeling Item-specific Effects for Video Click, SIAM International Conference on Data Mining, San Diego, USA, May 2018. (Acceptance Rate: 87/375=23.2%)

F. Tan, C. Cheng, and Z. Wei, Time-aware Latent Hierarchical Model for Predicting House Prices, 17^{th} IEEE International Conference on Data Mining, New Orleans, USA, November 2017 (Acceptance Rate: 155/778=19.9%)

F. Tan, C. Cheng, and Z. Wei, Modeling Real Estate for School District Identification, 16^{th} IEEE International Conference on Data Mining, Barcelona, Spain, December 2016 (Acceptance Rate: 178/904=19.6%)

W. Zhang, Y. Xia, and **F. Tan**, Oscillations in Interconnected Complex Networks under Intentional Attack, International Journal of Modern Physics C, DOI: 10.1142/S0129183116500595, 2016 ($\mathrm{IF}_{2016} = 1.171$)

S. Banerjee, Z. Wei, **F. Tan**, K. Peck, N. Shih, M. Feldman, T. Rebbeck, J. Alwine, and E. Robertson, Distinct Microbiological Signatures Associated with Triple Negative Breast Cancer , Sceintific Reports, DOI: 10.1038/srep15162, 2015 ($\mathrm{IF}_{2016} = 4.259$)

F. Tan, Y. Xia, and Z. Wei, Robust-Yet-Fragile Nature of Interdependent Networks, Physical Review E, DOI:10.1103/PhysRevE.91.052809, 2015 ($\mathrm{IF}_{2016} = 2.366$)

F. Tan, Y. Xia, and B. Zhu, Link Prediction in Complex Networks: A Mutual Information Perspective, PLoS ONE, DOI: 10.1371/journal.pone.0107056, 2014 ($\mathrm{IF}_{2016} = 2.806$)

F. Tan, J. Wu, Y. Xia, and C. K. Tse, Traffic Congestion in Interconnected Complex Networks, Physical Review E, DOI: 10.1103/PhysRevE.89.062813, 2014 ($\mathrm{IF}_{2016} = 2.366$)

F. Tan, Y. Xia, W. Zhang, and X. Jin, Cascading Failures of Loads in Interconnected Networks under Intentional Attack, Europhysics Letters, DOI: 10.1209/0295-5075/102/28009, 2013 ($\mathrm{IF}_{2016} = 1.957$)

F. Tan, Y. Xia, Hybrid Routing on Scale-Free Networks, Physica A, DOI: 10.1016/j.physa.2013.04.032, 2013 ($\mathrm{IF}_{2016} = 2.243$)

To My Beloved Ones

ACKNOWLEDGMENT

First and foremost, I wish to take this opportunity to express my heartfelt appreciation to my doctoral advisor Dr. Zhi Wei. It is his generous help and sustained encouragement that bring me the courage to overcome the challenges in the road of pursuing my dream and hunting for the scientific truth. Dr. Wei is my friend and mentor. It is his tremendous effort, invaluable guidance, and infinite patience that enable me to bring this dissertation to its culmination. I will always be indebted to Dr. Wei for generously supporting and inspiring me in this critical stage of my life.

Second, I am extremely grateful to Dr. James M. Calvin, Dr. Senjuti Basu Roy, Dr. NhatHai Phan and Dr. Wenge Guo for serving on my committee. They have provided me with academic advice inside and outside my research field. This dissertation would not have been possible without their invaluable guidance and generous help. In addition, I would like to extend special thanks to my collaborators, Dr. William Yan and Dr. Jun He from Adobe Inc., Dr. Yifan Hu, Dr. Changwei Hu and Dr. Aasish Pappu from Yahoo! Research. I wish to thank Dr. Ali Mili, Dr. Reza Curtmola, Dr. Cristian M. Borcea, Dr. George Olsen, Ms. Clarisa Gonzalez-Lenahan and Dr. Yongxiang Xia for supporting and helping me all the time. I would also like to thank my lab mates and all graduate colleagues for their assistance and support.

Third, I truly appreciate my family for being my emotional anchor. I thank my beloved parents and fiancee (M.D., Ph.D.) for their endless love, support, and encouragement. Lastly, but not the least, I am thankful to all who have helped me

directly or indirectly for the past five years. It is their support and encouragement that give me the courage and strength to pursue my PhD degree abroad.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF TABLES (Continued)

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure Page

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

BACKGROUND

Spatial and temporal data usually refer to data of space and time measurements. The presence of spatial and temporal information enables the new paradigms for analyzing data. The rich variety of problems with spatial and temporal dimensions cultivate the field of spatiotemporal data mining [6]. It has been widely explored in different applications including climate and environment (e.g., global change) [164], urban computing (e.g., land-use classification) [174], epidemiology (e.g., monitoring and predicting spread of disease) [112], criminology (e.g., crime hot spots) [29], mobile-commerce (e.g., location-based services) [100]. In the analytics of traditional data without spatial or temporal considerations, data are assumed to be independently generated. This presumption, however, doesn't hold true for spatial and temporal data due to the complexity of the types and relationships of data samples, features and target variables. Due to the interdisciplinary nature of spatiotemporal data mining [143], this dissertation covers the spatiotemporal representations in the context of involved application domains from three perspectives. They are *sample-wise* or *instance-wise*, *feature-wise* and *target-wise* dependencies, respectively.

More specifically, sample-wise dependency refers to that data instances are correlated with each other with varying properties in different spatial regions or time periods. This is the most widely perceived spatiotemporal dependency. For example, in urban computing, neighborhoods as a whole are likely to have some dependence among different houses based on the geolocation. In this section, we mainly study school district identification, housing price and video click prediction. They are confronted with heterogeneous information integration, the insufficiency of historical data and sparsity of learning features, respectively. For school district appraisal, the most commonly used methods are based on the relationships between sales prices of houses and school quality [46, 74]. However, school quality is not the sole determinant of the housing price, which is impacted by other housing attributes as well. Disentangling the effect of school and then inferring the school quality is a direction to correct for the bias caused by the sole relationship. In addition, integrating the geographical dependence among different houses into the multiple relationships is also a challenging task. For the housing price prediction, the classical hedonic pricing approach works well when all housing characteristics are available [118], which is not feasible in practice. To this end, Latent Manifold Estimation models the overall external characteristics of a house as a static latent variable by considering neighboring houses. However, no temporal interactions are considered due to the insufficiency of sales [34]. Some repeat sales and autoregressive models are also developed to track market trends by utilizing homes sold multiple times [65, 120]. These methods are based on a assumption that no significant housing features change between sales and the multiplicity of sales for same houses exist. These two requirements are self-contradictory and thus definitely limit their applications. For repeated video click records with sparse features, some regression based methods are proposed to construct pairwise feature space to scale up the predictive performance [129]. However, the performance gain of augmented features are subject to the quality of sparse features themselves and independence of records are presumed as well. In practice, repeated click records on same videos have temporal interactions to some degree, which is yet to explore to alleviate the feature sparsity.

The second section is the feature-wise dependency, which refers to spatial or temporal interactions among different internal features within same data instances. For example, subscribed users' daily activities have some temporal interactions, whereas different users have no explicit connections. We mainly explore two applied domains: user intended action prediction and DNA methylation. They are faced with heterogeneous feature integration and homogenous sequence representation, respectively. Specifically, regarding user intended action prediction (e.g., attrition forecasting), there are user activities logs, dynamic and static user profiles involved in this problem. Classical modeling strategies approach this problem either by concatenating all features directly without preserving the temporal constraints among them [36,171] or by capturing the temporal correlations with handcrafted efforts [148]. The potential of temporal interactions on the forecasting goal is not taken advantage of fully. Regarding DNA methylation prediction and interpretation, the spatial dependency among homogenous nucleotides usually forms the gene regulatory motifs, which are closely related to DNA methylation. The conventional machine learning approaches [84, 93] to genomic sequence-based prediction are based on the human handcrafted *k-mer* scheme[1]. However, they can only capture a limited spectrum of flanking context sequences. For motif elucidation, conventional motif analysis assumes simple independence and additive effects among regulatory bases [75]. It extracts less

[1] https://en.wikipedia.org/wiki/K-mer

sophisticated regulatory patterns, which accounts for only a small proportion of all methylated sites.

The last section is target-wise dependency. It means that dependent variables of interest are correlated with each other spatially and temporally, whereas instances have no explicit corresponding connections. For example, loans usually have different survival time or payment lifespans. The target variable survival time is perceived to be temporally correlated here. In this section, we focus on the competing risks representation in peer-to-peer lending. The principal issue is the characterization of both qualitative status and quantitative survival time for loans. The common strategy is to categorize loans into two simple statuses without considering the survival time [22, 183]. The classical survival analysis with competing risks focuses on the way multiple risks shape survival time jointly [111]. This cannot care for investors' concerns of both survival time and the underlying causing events fully.

In this dissertation, we study the above three genres and develop a set of methodologies to address heterogeneous features integration, homogenous sequence representation, the insufficiency of historical instances, the sparsity of features and unified characterization of status and time involved in spatiotemporal data mining for the aforementioned divergent domains. In the following part, we detail the challenging issues of specific application domains and brief our approaches in the order of sample-wise, feature-wise and target-wise dependencies. Overall, the proper representation strategies make it possible to unleash the potential of spatiotemporal interactions for the forecasting tasks and obtaining in-depth insights.

1.1 Sample-wise Dependency

1.1.1 Modeling Real Estate for School District Identification

The importance of a desirable educational environment to the choice of neighborhoods can never be overemphasized around the world. According to the "2010 Profile of Home Buyers and Sellers" from the National Association of Realtors (NAR)[2], more than half of all homebuyers with children under 18 years of age rate the quality of the local school district as a major factor influencing their choice of a neighborhood[3]. Such an observation also holds true for Chinese homebuyers, which can be illustrated by one of the most famous traditional Chinese idiomatic allusions, "*Mencius's mother, three moves*". It refers to the legend that Mencius's mother changed her residence three times on account of her concern for Mencius' education[4]. The school-district not only is closely related to education service itself but also can be regarded as a worthwhile investment. It's reasonable for parents to secure a stellar education for their children through buying homes associated with top schools. Even for the homebuyers who want to invest on real estate, those houses at excellent school districts might protect them from the market's ups and downs[5].

Inspired by these insights, the following interesting and practical question emerges: how to identify neighborhoods with high-quality school services automat-

[2] http://www.mdrealtor.org/Portals/0/docs/ResearchandStatistics/PROFILE%20OF%20BUYERS-SELLERS%202010.pdf

[3] a geographically localized community within a larger city, town, suburb or rural area. In most urban areas of Mainland China, neighborhood usually refers to the residential community or unit grouped by multiple families, and it is the direct sub-level of a subdistrict, which is the direct sub-level of a district, which is the direct sub-level of a city.

[4] https://en.wikipedia.org/wiki/Mencius

[5] http://www.sfgate.com/education/article/BAY-AREA-NEW-MEANING-TO-API-SCORES-HOME-2567517.php

ically? This might be easy for American estate market due to both well-defined association between neighborhoods and school districts[6] and comprehensive school ranking system[7]. However, it still remains a crucial missing chapter in the real estate appraisal field for other countries with unbalanced education resources and poor information disclosure. Taking China for instance, there are no official or systematic ratings of primary schools. In this case, the decision-making procedure is actually directed by either word-of-mouth information or online empirical comments. The procedure, however, is highly time consuming due to numbers of factors involved. Therefore, the issue of the school-district of a home is more complicated for this kind of real estate market.

Before tackling this question, the challenging parts are elaborated as follows: Firstly, how to disentangle the effect of the educational environment on local real estate pricing without any prior knowledge about the quality of the school district. An abundance of studies have reported the interplay between residential neighborhoods and associated local school services [37, 63, 73, 80, 88]. The general point is that educational services are widely believed to be a key determinant of housing prices [37,88]. However, many other factors like housing type, building year and surrounding traffic conditions impact the property value as well. Secondly, how to assess the quality of school services based on raw textual comments. Besides the housing value, some online textual comments on neighborhoods can provide us with clues to tell excellent school districts from mediocre ones. Nonetheless, apart from education, the reviews normally involve many other facets of the neighborhood. Thirdly, how

[6] http://schooldistrictfinder.com
[7] http://www.schooldigger.com

to integrate geographical dependence between adjacent neighborhoods into previous two explorations. Intuitively speaking, adjacent neighborhoods are more likely to share similar educational service quality than remote ones. However, the degree of similarity amongst nearby neighborhoods is latent and thus needs to be figured out.

To this end, we propose a geography-based latent variable hierarchical probabilistic framework. The aforementioned challenges can thus get addressed in an elegant manner. Regarding housing prices, a latent linear regression model is developed to infer the impact of school districts. However, this is insufficient due to the limited availability of attributes and complexity of price dynamics [159]. To overcome this deficiency, we employ topic model [18] to extract education-related topics from raw comments and then develop a multinomial mixture model. They are combined together and learn mutually reinforced knowledge from both numerical data and textual information to capture the intrinsic educational impact of neighborhoods. Furthermore, geographical dependence among neighborhoods is modeled as a discrete local Markov Random Field (MRF) and explicitly incorporated as *a priori* into the hierarchical probabilistic framework.

1.1.2 Modeling and Elucidation of Housing Price

Many online real estate database companies such as *Zillow*, *Realtor*, *Redfin* and *Lianjia*[8] provide functions to estimate the market value of an individual home. An appealing estimation performance can produce a good starting point in determining house prices for homebuyers. The improvement of customers' loyalty to the websites,

[8] *http://bj.lianjia.com*

in turn, can bring about the increase of the potential revenue. Furthermore, housing market dynamics are closely related to local economic prosperity. The evolution of real estate values is capable of providing reasonable insights in this regard. Therefore, the accurate prediction of house prices is informative and beneficial for homebuyers, online platforms and even economic study.

Essentially, housing prices vary spatially from municipalities, local communities to houses themselves and temporally from one transaction period to another one. The variation can be attributed to geographical location based socioeconomic conditions (e.g., environment, education, income levels, population density, and demographic effects), internal housing characteristics (e.g., lot size, square footage, and number of rooms), and temporal effects (e.g., governmental regularization policy, economic development, and marketplace of demand and supply) [40]. Therefore, the task of housing price prediction aims to estimate the future market value of houses, given the precedent transaction records of houses (but not necessarily same houses). This problem, however, is facing several challenges.

First, it is traditionally difficult to collect all housing characteristics on a broad scale. Each house has a large number of attributes such as location, lot size, square footage, number of bedrooms and living rooms and many other details. The availability of housing characteristics depends on the extent to which they are released by property agents and house owners. Second, it is tricky to quantify some location-associated socioeconomic characteristics. For example, there are no unified metrics to measure the overall quality of the environment and education resources. However, those attributes might play a crucial role in the evolution of property value.

Furthermore, those socioeconomic features are dynamic rather than constant. Third, the real-world market value of a home can only be observed when the transaction activity does occur. Different from a regular product, a house, however, is less likely to be frequently sold within a short time span. Thus, the transaction records of same houses are usually sparse within a typical study time span of interest.

Formally, house prices are assumed to be impacted by both individual housing and common socioeconomic features in this dissertation. Obviously, individual characteristics vary across different houses, which are typically internal features. On the contrary, socioeconomic features are usually associated with location and are shared by different houses within a neighborhood to some extent. They are usually external features including but not limited to environment, education, transit facilities, living facilities, income level. Put it another way, we roughly regard individual housing and common socioeconomic features as *internality* and *externality* respectively as detailed in Table 3.3. The aggregate effects of corresponding groups on the housing price are called internal and external components, respectively. Particularly, the higher the value of an external component is, the more desirable the corresponding location is.

Following this perspective of the housing price and aforementioned challenges, two questions are emerging accordingly. Is it possible to infer a surrogate of the external component from historical transaction records by only leveraging location information? What roles do the external and internal components play in housing price dynamics, respectively (The answer will be given in Section 3.4.3)? Therefore, we propose the concept of *neighborhood value* as the above surrogate, which is associated with a specific neighborhood. Although the neighborhood value is abstract and latent, it has the underlying structural mechanism behind the observed house prices. To be specific, in spatial dimension, neighborhood value doesn't change sharply among nearby neighborhoods in general. This assumption also applies to the temporal dimension. That is to say, the value changes gradually from one time period to the successive one [186]. Such spatial and temporal smoothness constraint aids in the inference of latent neighborhood value. The sparsity of transaction records of same houses during a short time period still exists, which is particularly faced by repeat sales methods [10, 120]. To address this issue, we group different houses into the same neighborhood based on the predetermined criterion. In our dataset, the neighborhood is a residential unit or quarter of 100 to 600 families distributed in several buildings[9]. Apparently, the selling prices of any houses within a neighborhood can provide essential information for the inference of latent neighborhood value. Thus, such a strategy obviates the requirement of the long time span of the transaction data. The basic idea of neighborhood value inference is to assign a time-dependent value to each neighborhood during different time periods when the training phase is implemented. The value of one neighborhood is obtained by aggregating both the weighted value of nearby neighborhoods and its precedent value. The time-aware latent hierarchical model is thus proposed to capture how the latent neighborhood value forms and evolves.

We conduct comprehensive experiments on a real-world real estate transaction dataset. The proposed model is demonstrated to achieve better performance over

[9] https://en.wikipedia.org/wiki/neighborhood

baseline algorithms. To explore the roles of both components in housing price prediction, we further perform hierarchical feature analysis. It's found that the external component is dominant in house prices, whereas the internal one only impacts them marginally. Furthermore, the inferred neighborhood value is experimentally shown to be capable of approximating the external component properly. That is to say, we can still achieve an appealing performance even if a large volume of location associated features are unavailable.

1.1.3 Modeling Item-specific Effects for Video Click

Video click is a crucial concern for video content providers. High video click rate is able to bring about more profits for video websites. Thus, to ensure as high a click rate as possible is one of their important commercial goals. To this end, a wealth of associated prediction schemes have been proposed in this regard.

These powerful strategies involve general-purpose algorithms [2, 31, 107, 129, 165] and specialized methods customized to videos [12, 30, 44, 61]. From the perspective of feature representation, the aforementioned algorithms mainly focus on exploring and modeling available features for prediction. Even though the existing methods can provide powerful feature representation and achieve appealing prediction performance, there are still room for further improvement. This is because it is very hard and/or prohibitively expensive to collect or record all raw features involved in click actions due to many practical difficulties and constraints. Put another way, there are generally some hidden features (uncollected features instead of features with missing value) for learning algorithms. In this case, we hypothesize there are some consistent effects in multiple click records of the same video. If hidden features contain such effects, one promising direction is to bring them back in order to alleviate the above issue.

To facilitate the understanding of the potential consistent effects, we here present a classical example of student academic performance. Roughly speaking, the academic performance of a student is dependent on a number of attributes including students' talent, parental support, school academic quality and so on. Specifically, students within the same classroom seem more likely to receive similar teaching and academic guidance than students in different classrooms. The classroom-level consistency among students is supposed to exist in all features related to the classroom irrespective of their availability and measurability. However, only partial features get gleaned for study in practice. Those hidden features probably involve the consistent effects associated with classrooms. There are two general approaches to rescue them. A simple one is to apply one-hot encoding scheme directly[10] to model classrooms as a categorical feature. In this case, the number of unknown parameters increases with the number of classrooms, which is inconsistent in parameter estimation [124]. A compromised approach is to roughly group classrooms into coarse-grained categories according to other available features. It yields consistent and effective parameter estimation at the cost of losing important fine-grained classroom information. As a matter of factor, the latter approach is implicitly adopted in the feature engineering of many studies.

[10] https://en.wikipedia.org/wiki/One-hot

Thus, in this dissertation, we introduce a tradeoff between these two alternatives to characterize such speculated effects for clicking records. To simplify the elucidation of the methodology, we take the regression-based algorithm [129, 165] as a basic model for study. To be concrete, we introduce a simple yet effective variable into the classical logistic regression to capture the hypothesized effects in hidden features. A series of thorough simulation studies demonstrate that the intrinsic effects can be rediscovered effectively if they indeed exist. Furthermore, the more salient such effects are, the more the current models get boosted in terms of predictive power. In addition, we conduct the comparison between the proposed algorithm and the existing counterparts on click records of a real-world video platform. The empirical results validate the existence of such grouped effects in clicking behaviors. Since the group of click records are clustered on the item of video, we also call them *item-specific effects*.

1.2 Feature-wise Dependency

1.2.1 Blended Learning for Predicting User Intended Actions

Being able to predict user intended actions and elucidate underlying behavior patterns are of significant value for the business development. Such intended actions include, but not limited to, user conversion (e.g., purchase, signup), attrition (e.g., churn, dropout), default (failure to pay credit cards or loans), etc. These user actions directly lead to revenue gain or loss for companies. The capability of predicting user intended actions may help companies to take proactive measures to optimize business outcome.

In this dissertation, we focus on predicting attrition, which is one of the most representative user intended actions. Attrition, in a broad context, refers to individuals or items moving out of a collective group over a specific time period[11]. It can be specialized, as seen in broad applications in different fields. For example, Massive Open Online Courses (MOOCs)[12] can offer an affordable and flexible way to deliver quality educational experiences on a new scale. However, the accompanied high dropout rates are a major concern for educational investors [70]. In the commercial context, the revenue growth of enterprises heavily relies on the acquisition of new customers and retention of existing ones. Previous researches and reports have shown that retaining valuable customers is cost effective and more rewarding than acquiring new customers [162, 169].

Accordingly, targeting at-risk attrited users in advance and taking intervention measures proactively is crucial for improving students' engagement and maintaining customers' retention. It helps to sustain the prosperity of MOOCs and enterprises.

There are, however, several inherent challenges confronted in predicting attrition using user usage data. (1) User alignment is a tricky problem as the improper alignment may incur intrinsic bias in the subsequent modeling; (2) Multi-view heterogeneous data sources, ranging from user activity logs to dynamic and static user profiles, pose a barrier to the effective interaction and amalgamation; (3) It is not a trivial task to characterize primitive user activity logs, let alone integrating them with the downstream predictive modeling effectively and seamlessly; (4) How to keep track of the evolving intentions of observed historical records for improving attrition within a target time period has yet to be explored fully; (5) It remains

[11] https://en.wikipedia.org/wiki/Churn_rate, https://www.ngdata.com/what-is-attrition-rate/

[12] http://mooc.org

unclear how to quantify and visualize the importance of underlying activity patterns, attrition and retention factors.

To address these challenges, we revisit the attrition problem from both predictive modeling and underlying patterns representation sides. To be specific, we first introduce an appropriate user alignment scheme based on the calendar timeline, which can remove the bias as mentioned before. Under an unbiased framework, we propose a *Blended Learning Approach* (BLA) to address related issues, which renders an appealing predictive performance. BLA is mainly characterized as multi-path learning, intention guidance and multi-snapshot mechanism. The multi-path learning embeds heterogeneous user activity logs, dynamic and static user information into an unified learning paradigm. The multi-snapshot mechanism integrates historical user actions explicitly into the model learning for tracking the evolution of patterns, which is further enhanced by the intention guidance and decay strategies. For multi-snapshot mechanism, the summarization strategy is developed to bridge the separation of the labor-intensive aggregation of user activities and model learning. The model performance is evaluated on two public data repositories and one dataset of Adobe Creative Cloud user subscriptions. Furthermore, a simple yet effective visualization approach is introduced to discover underlying patterns and to identify attrition and retention factors from user activities and profiles. This may be exploited by the business or educational units to develop a personalized retention strategy for retaining their users.

1.2.2 Elucidation of DNA Methylation on N^6-Adenine

DNA methylation is extensively involved in epigenetic settings and exerts different regulatory roles in multiple species [78, 106, 176]. It is traditionally acknowledged that 5-methylcytosine (5mC) presents a dominant modification in eukaryotes, while N6-methyladenine (6mA) is mostly prevalent in prokaryotes [51]. Thanks to the development of high-throughput sequencing (6mA-IP-seq) and single-molecule real-time (SMRT) sequencing technology, the prevalence and significance of DNA 6mA in eukaryotes (e.g., A. thaliana and D. melanogaster) have been revealed recently [60, 68, 102, 170, 179]. However, DNA 6mA is a dynamic process, which can be developmental and tissue specific [106]. In addition, many 6mA sites may be methylated at very low levels, making them very hard to be captured. Consequently, current experimental approaches, although precise, are unable to provide a complete catalog of all 6mA sites. It has been long recognized that 6mA plays a vital role in discrimination of host genomic DNA from foreign pathogenic DNA in bacteria [106, 178]. Recently, it has been demonstrated that 6mA may be involved in gene activation or repression in eukaryotes [59, 178]. The underlying mechanism, however, remains elusive. Methylation-associated gene regulatory motifs may shed light on understanding the mechanism. Although conventional motif analysis of 6mA has revealed some interesting cis-regulatory patterns, they account for only a small proportion of all methylated sites [75, 106]. Therefore, we hypothesize that more sophisticated regulatory mechanisms yet to be explored may exist for 6mA formulation. Lastly, the whole in vivo cataloguing procedure of 6mA is costly and laborious. Thus, in silico prediction may be an attractive alternative if we can precisely predict 6mA sites at single-nucleotide resolution based on just genomic sequence information.

1.3 Target-wise Dependency

1.3.1 Competing Risks Representation in Peer-to-Peer Lending

Peer-to-Peer (P2P) lending has become a fast-growing new channel of financing over the past decade. Quite a few P2P platforms have been developed including Lending Club[13], Prosper[14], Yirendai[15] and Zopa[16]. Connecting borrowers with investors directly using technology, those P2P platforms claim to operate at a lower cost than traditional bank loan programs, passing the savings on to borrowers in the form of lower rates and to investors in the form of solid returns. Such credit marketplaces have thus attracted a lot of lenders (investors) and borrowers and result into a large amount of investments. For example, as of June 30, 2018, the total loan issued by Lending Club (the world's largest P2P lending platform) has exceeded 38 billion US Dollars[17].

For P2P lending, three major participants are involved in the transaction procedure: the lending platform, lenders and borrowers. Lenders and borrowers interact with each other directly on the lending platform. Using Lending Club for example, we briefly introduce the working mechanism of P2P lending. Other P2P lending platforms are somewhat similar. In Lending Club, a borrower (sometimes with

[13] www.lendingclub.com

[14] www.prosper.com

[15] https://www.yirendai.com

[16] www.zopa.com

[17] https://www.lendingclub.com/info/statistics.action

co-borrowers) is supposed to provide his or her detailed profile (e.g., annual income, housing status) and loan information while creating a listing to solicit investments from lenders. After receiving the listing, the platform verifies borrowers' profile (optional), evaluates their credit, and then assigns a certain grade or sub-grade to the listed loan for lenders' reference. If the listed loan gets fully funded by the expiration date, it will be issued by the platform, or otherwise revoked. Afterwards, the investors can secure interests and the platform charges service fees from borrowers' monthly payments. Like in most conventional bank loan programs, borrowers may prepay their loans at any time, in whole or in part, without penalty; lenders will then receive pro rata share of the payment. A loan can also become charged off when there is no longer a reasonable expectation of further payments.

Several P2P lending platforms release their loan data to the public, which has received much attention from academia [22, 43, 47, 101, 105, 182, 183]. The existing works mainly involve simple binary classification between types of charge-off and full-payment [22, 47], loan recommendation based on charge-off risk [183] and multi-objective portfolio optimization [182]. With regards to loan risk modeling, the common focus of previous works is on the overall charge-off risk.

However, the risk of prepayment (the settlement of entire balance of a loan before its official due date) is often ignored in online P2P lending study although prepayment has been well-studied in classical literature in other financial industries like mortgage risks [33, 114, 138, 146]. As with charge-off, prepayment would also terminate the repayment schedule. In this case, charge-off and prepayment are two competing risks as they coexist in the same loan over the course of loan repayment.

Classical survival analysis with competing risks can be naturally utilized to model the risks of charge-off and prepayment for time-to-event loan data [111]. Nonetheless, their focus is on how survival time as a dependent variable is shaped by multiple risks jointly. This works well for many applications such as clinical trials and insurances, in which the time to event is the major concern, but not the underlying causes/risks. Different risks are just modeled as covariates simultaneously for a better estimate of the survival time. For P2P lending, however, investors are concerned with both survival time and the underlying causing events. There are three points beyond the focus of classical survival analysis with competing risks. (1) Under many circumstances, the latter might play a dominant role in investment performance. Consider, for instance, a loan with survival time of 5 months, for which the event of charged-off causes loss to investors, whereas the state of prepayment leads to positive returns. It is thus necessary to distinguish different events properly. This also largely explains why the previous research efforts focus on coarse-grained binary statuses. (2) Meanwhile, the earlier the prepayment (charge-off) occurs, the less preferable a loan would be. To put it another way, the discrete survival time of a loan is inherently ordinal for the same risk. (3) What's more, the eventual events cannot co-occur, and the same event cannot occur multiple times for the same loan either. They are actually exclusive with each other irrespective of events or the survival time. Therefore, we propose to model the coarse-grained rivalry between different events, fine-grained competition of survival time within the same event and underlying ordinal constraints explicitly and simultaneously.

To this end, we first propose a grading rule for each risk independently on the basis of the survival time. We then transform the hierarchical ordinal regression problem to multiple binary classification sub-problems [32, 95]. Under the newly formulated framework, we further integrate censored loans without definite observed events into the representation. A hierarchical fine-grained risk categories with ordinal constraints can be generated accordingly. Simultaneously, we fuse loans of multiple scheduled terms by introducing a masking layer. An architecture of deep neural networks with multiple risk category outputs of multiple terms is finally proposed.

CHAPTER 2

MODELING REAL ESTATE FOR SCHOOL DISTRICT IDENTIFICATION

2.1 Introduction

Overall, real estate appraisal has been covered by a large volume of research from various perspectives. The affiliated school district of a real estate property is often a crucial concern, especially for those homebuyers with school-age children. How to evaluate the quality of school districts properly and automate the identification of residential homes located in a favorable educational environment, however, is largely unexplored until now. The availability of heterogeneous estate-related data offers great new opportunities for harnessing the power of diversified information correlated with education for school district assessment. Nevertheless, it is such heterogeneity that poses a significant challenge to their amalgamation for identification in a unified fashion. To this end, we develop a geographical latent variable hierarchical probabilistic model to integrate digital price, textual comments, and geographical location information together to assess residential environment in terms of its affiliated school- related services. The proposed approach is able to capture the in- depth interaction among multi-type data greatly. Our framework is further examined on the dataset of Beijing property market. The results justify the benefits of our approach over baseline methods. The further comparison among different components of the proposed model is also conducted and demonstrates their important role.

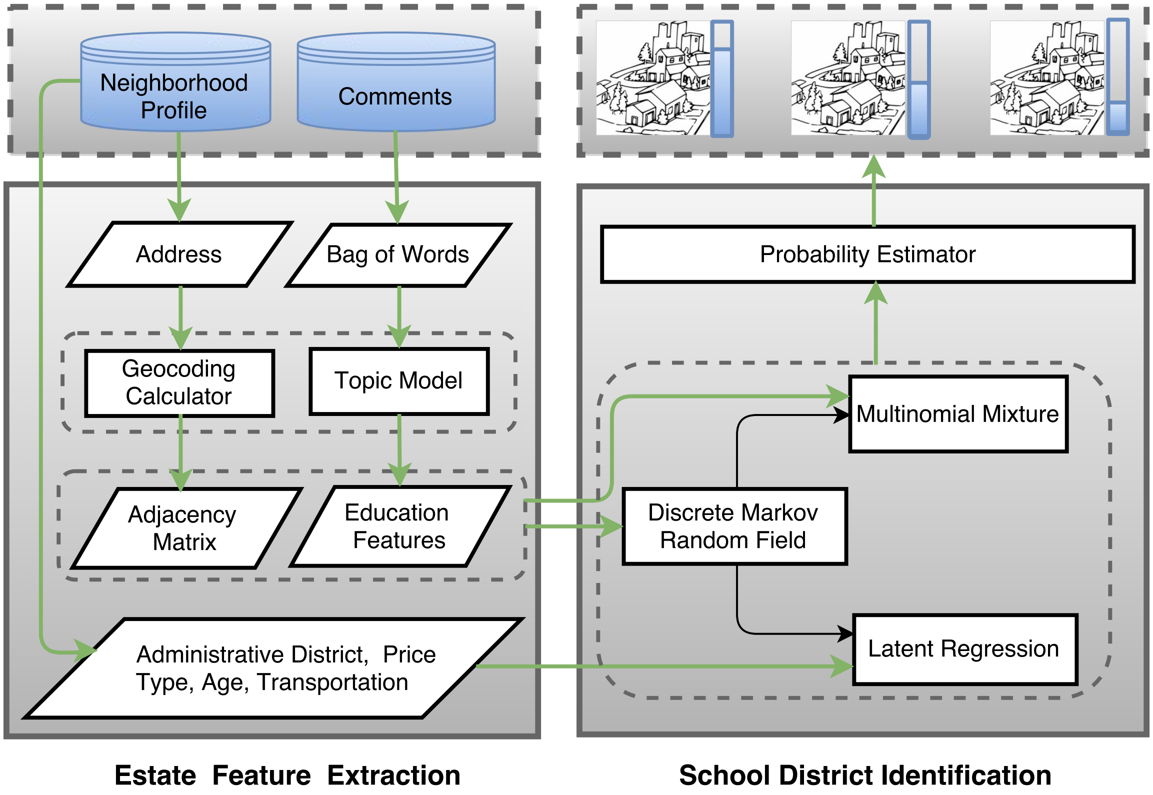


Figure 2.1 Framework overview of the school district identification.

Moreover, the proposed model is flexible and can offer useful insights into modeling heterogeneous data sources.

2.2 Related Work

Real estate appraisal: Previous research mainly focused on real estate appraisal itself [55–57, 120]. Clustering and ranking are performed simultaneously to predict estate investment values [57]. Fu *et al.* also conducted pairwise estate ranking by capitalizing on mobility behaviors and features extracted from user comments [55]. They continued to augment the performance of real estate appraisal by taking diverse mixed land use into account [56]. There is also some work on the relationship between sale prices and school quality, which is explored by building various indices and models [46,74]. These efforts, however, concentrated on detailed attributes of schools. In contrast, our goal is to identify quality school districts. To our best knowledge, this makes the first attempt.

Urban computing: Our work also relates to some topics in urban computing like POI recommendation [97, 173, 184]. For instance, Topic Model was applied to discover regions of different functions based on POIs and human mobility records [173]. They mainly examine the overall POI recommendation or functional region discovery. However, we focus on the interaction between residential and educational regions.

Hierarchical probability model: The proposed model is related to some classical mixture models [17,18,115] and MRF-based approaches [16,67,96,168]. For example, Blei *et al.* investigated different mixture models to boost the performance of image caption [17], in which the Gaussian-Multinomial Mixture model is a simplified version of latent regression analysis in this paper. These studies actually focused on the coarse-grained topics. Our focus, however, is primarily on fine-grained category of a specific topic.

2.3 School District Identification

2.3.1 Preliminary

If school quality is simply assumed to be good and ordinary, this problem can be formulated to estimate the probability with which a neighborhood is associated with a good school district. Formally, let $\boldsymbol{z} = \{z_1, z_2, \ldots, z_M\}$ be a set of state of M neighborhoods, each of which z_d is 1 if it has a good one and 0 otherwise. Essentially, our model is thus to estimate the probability of $z_d = 1$ in a unified way. The common symbolic notation rule is obeyed throughout this article: capital, lower-case bold letters respectively denote matrices and column vectors while non-bold letters represent scalars.

Regarding housing prices, (1) the collection of the unit housing price for M neighborhoods is represented as $\boldsymbol{y} = \{y_1, y_2, \ldots, y_M\}$; (2) The associated feature space is denoted as $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$. The d^{th} neighborhood has $\boldsymbol{x}_d = \{z_d, x_{d2}, \ldots, x_{dp}\}$. Here x_{d1} is replaced with hidden variable z_d to facilitate the notation of school district. For comments, (1) a corpus of comments on all neighborhoods is represented as $\boldsymbol{D} = \{\boldsymbol{w_1, w_2, \ldots, w_M}\}$; (2) words are supposed to be drawn from vocabulary indexed by $\{1, ..., T\}$ and document d associated with a neighborhood is a collection of N_d words denoted by $\boldsymbol{w}_d = \{w_{d1}, w_{d2}, \ldots, w_{dT}\}$ where w_{dt} is the number of the t^{th} item of vocabulary, thus $\sum_{t=1}^{T} w_{dt} = N_d$.

To facilitate the understanding of our framework, we build a set of increasingly sophisticated models, culminating in the *Geographical Latent Regression Multinomial Mixture model.*

2.3.2 Latent Regression-Multinomial Mixture

Model In this chapter, the quality of the affiliated school district is unknown, a *Latent Regression* (LR) model is thus developed [158]. Concretely, the unit price y_d

can be formulated as follows:

$$y_d = \alpha_0 + \alpha_1 z_d + \sum_{i=2}^{p} \alpha_i x_{di} + \varepsilon_d, d = 1, 2, \ldots, M \tag{2.1}$$

where error variable ε_d is assumed to be $\varepsilon_d \sim \mathcal{N}(0, \sigma^2)$ and independent of covariates. The coefficients are denoted as $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1, \ldots, \alpha_p\}$ for brevity. We rewrite Equation 2.1 as $y_d \sim \mathcal{N}(\mu_d, \sigma^2)$ where $\mu_d = \alpha_0 + \alpha_1 z_d + \sum_{i=2}^{p} \alpha_i x_{di}$ is the expectation and σ is the standard deviation. We explicitly have

$$y_d \sim \begin{cases} \mathcal{N}(\mu_{d1}, \sigma^2), & \text{if } z_d = 1 \\ \mathcal{N}(\mu_{d0}, \sigma^2), & \text{otherwise} \end{cases} \tag{2.2}$$

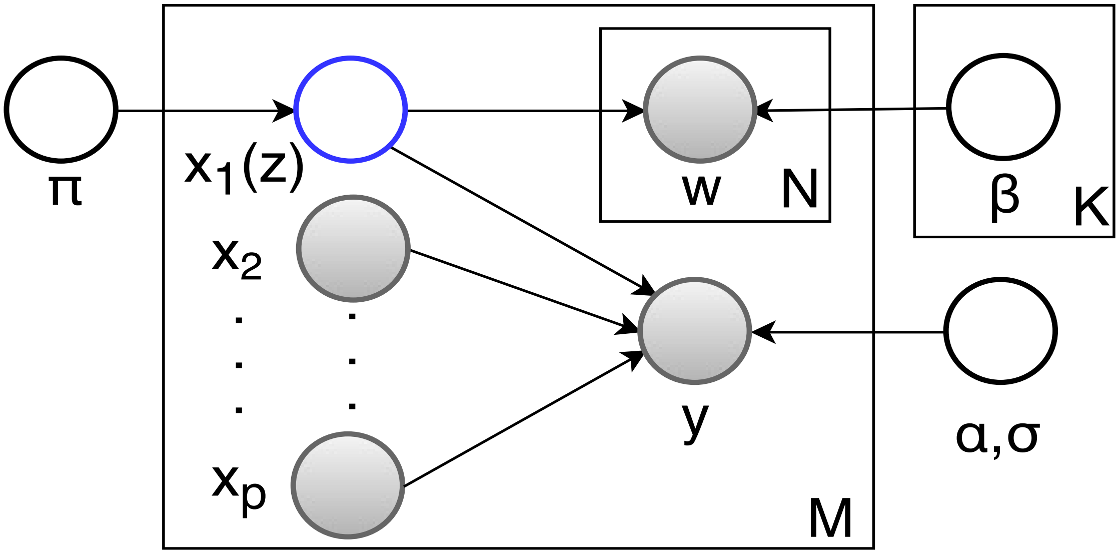


Figure 2.2 The graphical model representation of LRMM model. Shaded nodes are observed random variables; unshaded nodes are latent random variables, particularly, the blue unshaded node is the variable of interest.

where $\mu_{d1} = \alpha_0 + \alpha_1 + \sum_{i=2}^{p} \alpha_i x_{di}$ and $\mu_{d0} = \alpha_0 + \sum_{i=2}^{p} \alpha_i x_{di}$.

The semantic words are usually assumed to follow a multinomial distribution [18] with parameter $\boldsymbol{\beta}$, i.e., $\boldsymbol{w}_d \sim \mathcal{M}ult(\boldsymbol{\beta})$. Likewise, the distribution can also be explicitly described as

$$\boldsymbol{w}_d \sim \begin{cases} \mathcal{M}ult(\boldsymbol{\beta}_1), & \text{if } z_d = 1 \\ \mathcal{M}ult(\boldsymbol{\beta}_0), & \text{otherwise} \end{cases} \tag{2.3}$$

where $\boldsymbol{\beta}_k$ is symbolized as $\boldsymbol{\beta}_k = \{\beta_{k1}, \beta_{k2}, \ldots, \beta_{kT}\}$, $k \in \{0, 1\}$. It is called *Multinomial Mixture* (MM) model.

Latent Regression Multinomial Mixture (LRMM) model is formed by integrating them together, as shown in Figure 2.2. We introduce $\boldsymbol{\theta} = \{\pi, \boldsymbol{\alpha}, \sigma, \boldsymbol{\beta}\}$ to specify LRMM aggregately. The joint distribution of the hidden variable z_d and price/comments pair $(y_d, \boldsymbol{w}_d)$ parameterized by $\boldsymbol{\theta}$ is defined by

$$p(z_d, y_d, \boldsymbol{w}_d|\boldsymbol{\theta}) = \pi_k \mathcal{N}(y_d|\mu_{dk}, \sigma) \mathcal{M}ult(\boldsymbol{w}_d|\boldsymbol{\beta}_k) \tag{2.4}$$

where $\pi_k = p(z_d = k)$ is the mixing proportion and $\sum_k \pi_k = 1$. The corresponding marginal probability for $(y_d, \boldsymbol{w}_d)$ is

$$p(y_d, \boldsymbol{w}_d|\boldsymbol{\theta}) = \sum_k \pi_k \mathcal{N}(y_d|\mu_{dk}, \sigma) \mathcal{M}ult(\boldsymbol{w}_d|\boldsymbol{\beta}_k) \tag{2.5}$$

Parameter Estimation With the formulated marginal probability, parameters $\boldsymbol{\theta}$ can be optimized based on Expectation Maximization (EM) procedure [41], which is commonly applied to the model with latent variables.

The probability that $(y_d, \boldsymbol{w}_d)$ are generated from component k is $p(z_d = k|y_d, \boldsymbol{w}_d, \boldsymbol{\theta})$. The posterior probability denoted as $\gamma_k(z_d)$ for simplicity can be inferred based on Bayes' rule as

$$\gamma_k(z_d) = \frac{\pi_k \mathcal{N}(y_d|\mu_{dk}, \sigma)\mathcal{M}ult(\boldsymbol{w}_d|\boldsymbol{\beta}_k)}{\sum_l \pi_l \mathcal{N}(y_d|\mu_{dl}, \sigma)\mathcal{M}ult(\boldsymbol{w}_d|\boldsymbol{\beta}_l))} \tag{2.6}$$

The latent $\boldsymbol{\theta}$ can be derived based on sufficient statistics including observed $(\boldsymbol{y}, \boldsymbol{D})$ and fixed $\gamma_k(z_d)$ by maximizing the corresponding log-likelihood:

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{D}) = \sum_{d=1}^{M} \ln \sum_k \pi_k \mathcal{N}(y_d|\mu_{dk}, \sigma)\mathcal{M}ult(\boldsymbol{w}_d|\boldsymbol{\beta}_k) \tag{2.7}$$

with respect to π_k subject to constraint $\sum_k \pi_k = 1$, $\boldsymbol{\alpha}$, σ^2 and $\boldsymbol{\beta}$ subject to $\sum_{t=1}^{T} \beta_{kt} = 1$.

Mixing Proportion π:

$$\pi_k = \frac{1}{M}\sum_{d=1}^{M} \gamma_k(z_d) \tag{2.8}$$

Regression parameters $\boldsymbol{\alpha}$:

$$\begin{aligned}
\alpha_0 &= \frac{1}{M}\sum_{d=1}^{M} \big[y_d - \gamma_1(z_d)\alpha_1 - \sum_{i=2}^{p} \alpha_i x_{di}\big] \\
\alpha_1 &= \frac{\sum_{d=1}^{M} \gamma_1(z_d)(y_d - \alpha_0 - \sum_{i=2}^{p} \alpha_i x_{di})}{\sum_{d=1}^{M} \gamma_1(z_d)} \\
\alpha_i &= \frac{\sum_{d=1}^{M} x_{di}\big[y_d - \alpha_0 - \gamma_1(z_d)\alpha_1 - \sum_{j \notin \{0,1,i\}} \alpha_j x_{dj}\big]}{\sum_{d=1}^{M} x_{di}^2}, \\
i &= 2, 3, \ldots, p
\end{aligned} \tag{2.9}$$

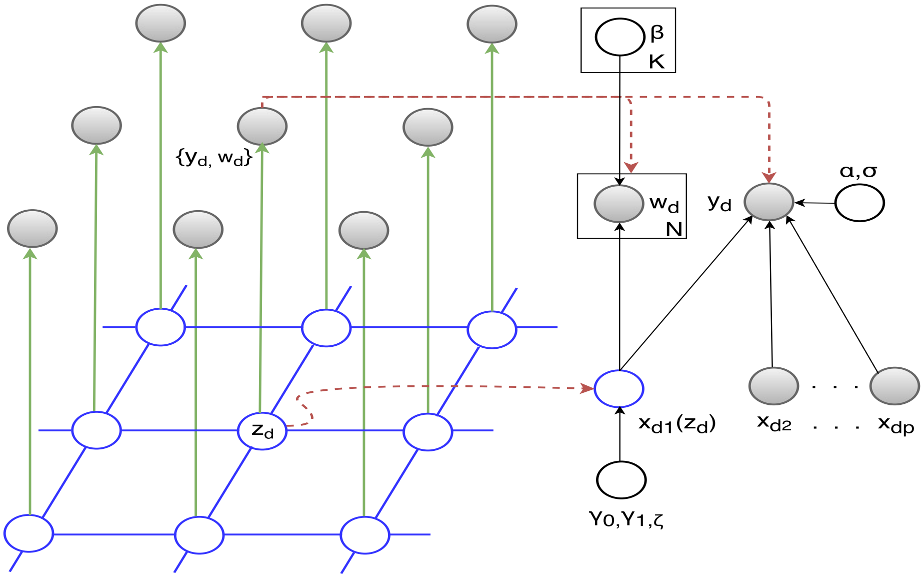


Figure 2.3 Graphical model representation of G-LRMM model.

Variance σ^2:

$$\begin{aligned}\sigma^2 = \frac{1}{M}\sum_{d=1}^{M}\Big[\gamma_0(z_d)(y_d - \alpha_0 - \sum_{i=2}^{p}\alpha_i x_{di})^2 \\ + \gamma_1(z_d)(y_d - \alpha_0 - \alpha_1 - \sum_{i=2}^{p}\alpha_i x_{di})^2\Big]\end{aligned} \tag{2.10}$$

Multinomial parameter $\boldsymbol{\beta}$:

$$\beta_{kt} = \frac{\sum_{d=1}^{M}\gamma_k(z_d)w_{dt}}{\sum_{d=1}^{M}\gamma_k(z_d)N_d} \tag{2.11}$$

2.3.3 Discrete Local MRF Model

Model The school district quality of different neighborhoods are actually not independent. For instance, if a neighborhood is associated with a good school district, the nearby neighborhoods are more likely to be located in favorable ones. To explicitly consider such dependence, we construct an undirected network $\boldsymbol{A}$ with the nodes for neighborhoods and edges for their connections. Here edges can be formulated by setting a threshold for the distance amongst all neighborhoods of interest. For M neighborhoods on the network, as stated earlier, $\boldsymbol{z} = \{z_1, z_2, \ldots, z_M\}$ is the vector of unobserved school district quality. In this dissertation, the dependence of $\boldsymbol{z}$ is thus modeled as an MRF with parameter $\boldsymbol{\Phi} = (\gamma_0, \gamma_1, \zeta)$. More specifically, the joint probability of $\boldsymbol{z}$ is assumed to be

$$p(\boldsymbol{z}|\boldsymbol{\Phi}) \propto \exp(\gamma_0 n_0 + \gamma_1 n_1 - \zeta n_{01}) \tag{2.12}$$

where $n_0 = \sum_d^M (1 - z_d)$ represents the number of neighborhoods at state 0, $n_1 = \sum_d^M z_d$ denotes the number of neighborhoods at state 1 and n_{01} is the number of edges connecting two neighborhoods with different states. Note we require parameter $\zeta > 0$ to discourage connected nearby neighborhoods to be in different states. For a specific neighborhood d, the probability of $z_d = k$ conditional on all others can be obtained by considering any two instances of all neighborhoods with different states only at neighborhood d. Concretely, we have conditional probability as

$$p_d(k|\cdot) \propto \exp(\gamma_k - \zeta\psi_d(1 - k)) \tag{2.13}$$

where $\psi_d(1 - k)$ represents the number of neighbors of neighborhood d with state $(1 - k), k = 0, 1$. We estimate parameter $\boldsymbol{\Phi}$ by maximizing the following conditional likelihood

$$\begin{aligned}\mathcal{L}(\mathbf{\Phi}; \boldsymbol{z}) &= \prod_{d=1}^{M} p(z_d|z_{\partial d}, \mathbf{\Phi}) \\ &= \prod_{d=1}^{M} \frac{\exp[(1-z_d)(\gamma_0 - \zeta\psi_d(1)) + z_d(\gamma_1 - \zeta\psi_d(0))]}{\exp[\gamma_0 - \zeta\psi_d(1)] + \exp[\gamma_1 - \zeta\psi_d(0)]}\end{aligned} \tag{2.14}$$

where $z_{\partial d}$ denotes the neighbors of neighborhood d.

We let $\boldsymbol{\eta} = (\boldsymbol{\alpha}, \sigma, \boldsymbol{\beta})$ be parameters for simplicity to specify the conditional probability $p(y_d, \boldsymbol{w}_d|z_d, \boldsymbol{\alpha}, \sigma, \boldsymbol{\beta})$ given state z_d. The condition can be denoted as

$$\begin{aligned}p(y_d, \boldsymbol{w}_d|z_d, \boldsymbol{\eta}) &= \left[\mathcal{N}(y_d|\mu_{d1}, \sigma)\mathcal{M}ult(\boldsymbol{w}_d|\boldsymbol{\beta}_1)\right]^{z_d} \\ &\times \left[\mathcal{N}(y|\mu_{d0}, \sigma)\mathcal{M}ult(\boldsymbol{w}_d|\boldsymbol{\beta}_0)\right]^{1-z_d}\end{aligned} \tag{2.15}$$

The log-likelihood of parameters $\boldsymbol{\eta}$ can be written as

$$\mathcal{L}(\boldsymbol{\eta}; \boldsymbol{y}, \boldsymbol{D}|\boldsymbol{z}) = \sum_{d=1}^{M} \ln p(y_d, \boldsymbol{w_d}|z_d, \boldsymbol{\eta}) \tag{2.16}$$

We place a geographical location-based MRF as *a priori* on LRMM and call such an integrated model G-LRMM for short. The overview of G-LRMM is shown in Figure 2.3.

Parameter Estimation The inference of true state $\boldsymbol{z}^*$ for M neighborhoods and parameter estimation must be carried out simultaneously. We propose the following iterative procedure based on iterated conditional modes (ICM) [16] to estimate $\boldsymbol{\eta}$ and $\mathbf{\Phi}$. The procedure is detailed in Algorithm 1.

Algorithm 1: G-LRMM

```
Input  : adjacency matrix $\boldsymbol{A}$, covariate matrix $\boldsymbol{X}$, price vector $\boldsymbol{y}$,
         education-related corpus $\boldsymbol{D}$
Output: an updated state vector $\boldsymbol{z}$
initialize states $\boldsymbol{z}^{(0)}$ and $\boldsymbol{\eta}^{(0)} = \{\boldsymbol{\alpha}^{(0)}, \sigma^{(0)}, \boldsymbol{\beta}^{(0)}\}$; initialize $\boldsymbol{\Phi}^{(0)} = (1, 1, 1)$;
repeat
    ## $\gamma_1(z_d) \leftarrow z_d^{(t)}, \gamma_0(z_d) \leftarrow (1 - z_d^{(t)})$;
    $\boldsymbol{\alpha}^{(t)} \leftarrow$ Equation (2.9);
    $\sigma^{(t)} \leftarrow$ Equation (2.10);
    $\boldsymbol{\beta}^{(t)} \leftarrow$ Equation (2.11);
    ## implemented by L-BFGS-B optimizer [23];
    $\boldsymbol{\Phi}^{(t)} = \underset{\boldsymbol{\Phi}}{\text{maximize}}\ \boldsymbol{\mathcal{L}}(\boldsymbol{\Phi}; \boldsymbol{z}^{(t)})$  # Equation 2.14;
    for $d \leftarrow 1, \ldots, M$ do
        $P(z_d^{(t+1)} | y_d, \boldsymbol{w_d}, z_{\partial d}^{(t)}) \propto p(y_d, \boldsymbol{w_d} | z_d^{(t+1)}, \boldsymbol{\eta}^{(t)}) p_d(z_d^{(t+1)} | z_{\partial d}^{(t)}, \boldsymbol{\Phi}^{(t)})$;
        if $P(z_d^{(t+1)} = 1 | \cdot) > P(z_d^{(t+1)} = 0 | \cdot)$ then
            $z_d^{(t+1)} = 1$;
        else
            $z_d^{(t+1)} = 0$;
        end
    end
    t = t + 1;
until Convergence;
return $\boldsymbol{z}$;
```

2.4 Experiment

In this section, we report empirical evaluation results of G-LRMM architecture on the real-world property market dataset.

2.4.1 Experimental Data and Preprocessing

We crawled the dataset of Beijing resale market from Lianjia[1] by September 18, 2015, a Chinese online real estate platform similar to Zillow.com. There are 17916 individual homes distributed among 1555 neighborhoods for sale in total, which received 48312 comments. In general, the eligibility of enrollment in specific elementary schools is similar for homes within the same neighborhood. Therefore, the raw data will be preprocessed at the neighborhood level.

Price Generally, housing prices are related to a large variety of characteristics. In our work, neighborhood-level ones are selected for our research including house types, transportation, administrative district, and building age. Transportation here refers to the availability of subway station within 1 km. Administrative district[2] refers to a subdivision of a city, which is a government-controlled sub-city.

Comments We use topic model to extract education-related topics from raw comments [18]. Specifically, we segment raw comments using Stanford Word Segmenter based on the Chinese Penn Treebank standard[3]. These comments are thus viewed as bags of terms (words) in vocabulary space. Then, LDA is applied to extract

[1] http://bj.lianjia.com

[2] https://en.wikipedia.org/wiki/District_(China)

[3] http://nlp.stanford.edu/software/segmenter.shtml

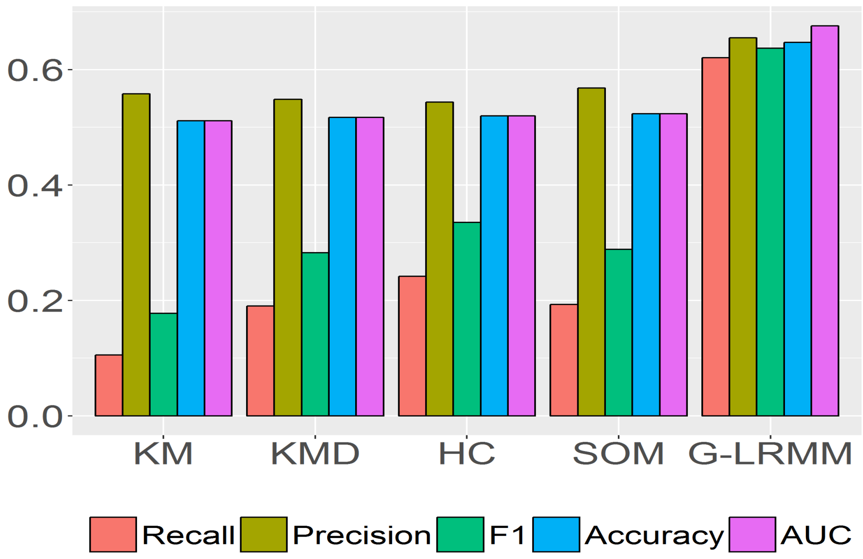


Figure 2.4 Performance comparison among G-LRMM model and clustering algorithms.

the potential educational topics[4]. The corresponding vocabulary can be defined based on top related terms. Finally, we construct a vector for each neighborhood.

Geographical Location We obtain geographical coordinate of each neighborhood[5] and then calculate the geodesic distance between any two neighborhoods[6], which is denoted as distance matrix. It is finally converted to adjacency matrix with a predefined distance threshold (500 meters (547 yards) as stated in Subsection 2.3.3 [97]).

[4] https://cran.r-project.org/web/packages/lda/index.html
[5] https://developers.google.com/
maps/documentation/geocoding/intro
[6] https://en.wikipedia.org/wiki/Haversine_formula

Ground Truth Data Since no systematic or official rankings for these elementary schools are available, we conduct the following studies to generate the ground-truth data. We collect the non-official partial ratings from different websites[7]. Those school districts receiving exactly same ratings from them are labeled as the common rating (good or ordinary). On the other hand, for school districts which receive different ratings or are not available in all databases, we resort to local people who have been in Beijing for over 7 years. Manually annotated labels are provided (good or ordinary).

2.4.2 Experimental Results

Evaluation Metrics The frequently used metrics are adopted: recall, precision, F1 score, accuracy and Area Under ROC Curve (AUC). Additionally, Precision@L and Recall@L are also used [56, 57, 97]. They are defined as Precision@L $= \frac{|z_L \cap z_P|}{L}$ and Recall@L $= \frac{|z_L \cap z_P|}{|z_P|}$, respectively. Here z_P is a set of neighborhoods with good schools and z_L is a list of top L ones sorted in the descending order given probability.

Baselines As our model is performed without labels, classical unsupervised learning algorithms are introduced as baseline schemes. (1) **KM**: K-means clustering based on Lloyd's algorithm ('kmeans' from R package 'stats' with parameter nstart = 100). (2) **KMD**: K-medoids clustering ('pam' from R package 'cluster'). (3) **HC**: Hierarchical clustering based on ward method ('hclust' from R package 'stats'). (4) **SOM**: Self-organising map-based hierarchical clustering ('somgrid' and 'som' from R package

[7] These databases include but not limited to https://geohey.com/data/public/education_school http://esf.fang.com/school/ http://bj.centanet.com/ershoufang-xuexiao/

‘kohonen’ with parameters xdim = 30, ydim = 30, rlen = 500). The input features are bag of words and price. The parameters are almost optimal.

To justify the necessity of complexity of G-LRMM, we also compare it with individual components. (1) **LR**: The latent regression model; (2) **MM**: Multinomial mixture model; (3) **LRMM**: It combines **LR** and **MM**; (4) **G-LR**: **LR** with geographical location dependence placed as the *priori.* (5) **G-MM**: **MM** with geographical dependence considered simultaneously. The parameters of models (1), (2) and (3) can be inferred via EM. Models (4) and (5) are optimized similarly to Algorithm 1.

Overall Performance and Analysis *Baselines*— The comparison results are shown in Figure 2.4. Overall, our model can present a significant performance gain on baselines. For recall, G-LRMM is over twice more likely than others to identify quality school districts. Such superiority of the proposed model is evident for both precision and F1 score. Aside from previous three metrics focusing on good districts, accuracy gives a bigger picture of the identification performance. As shown in Figure 2.4, the dominance still holds true for overall accuracy. Lastly, we calculate AUC to capture performance of different methods when the discrimination threshold is varied. The proposed model achieves the best performance.

Components— We explore the impact of different components of our method as detailed in Table 2.1. There are some observations. The first observation is that all methods with geographical information outplay the counterparts excluding locations significantly. Another one is that LRMM can achieve better performance than models LR and MM for almost all metrics. Such disparity between them mainly rises from

Table 2.1 Performance for G-LRMM and Its Components

Model	Recall	Precision	F1 Score	Accuracy	AUC
LR	0.0232	0.5806	0.0446	0.5035	0.6289
G-LR	0.5959	0.6325	0.6137	0.6251	0.6450
MM	0.4054	0.5422	0.4639	0.5318	0.5125
G-MM	0.5573	**0.6766**	0.6112	0.6457	0.6648
LRMM	0.5290	0.5948	0.5599	0.5846	0.5762
G-LRMM	**0.6203**	0.6549	**0.6371**	**0.6469**	**0.6756**

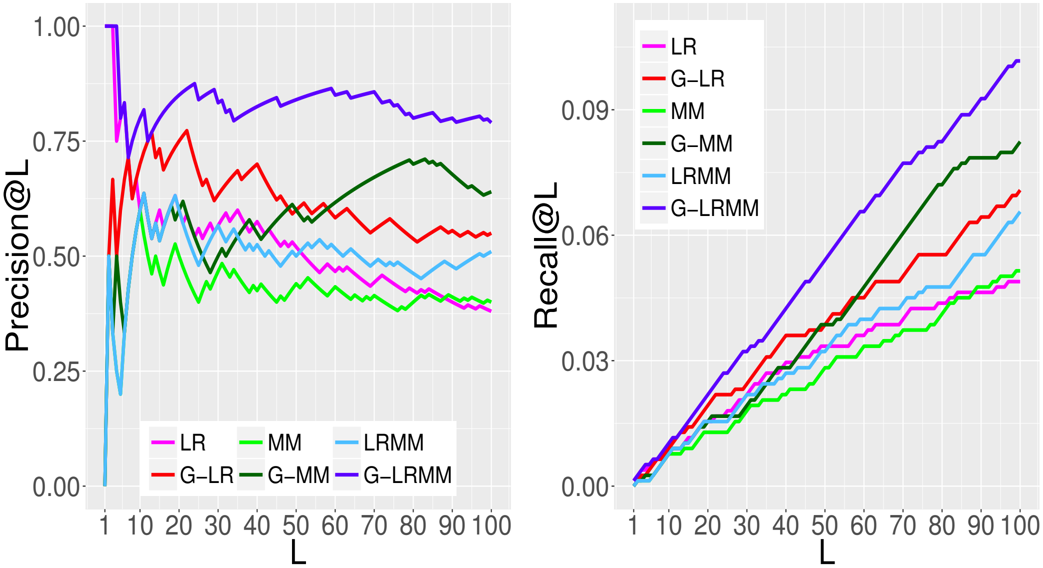


Figure 2.5 Precision and recall over top L neighborhoods.

the complementary interaction between price and comments. Furthermore, G-LRMM beats other two geography-based methods greatly. One exception is the slight edge of G-MM over G-LRMM in terms of precision. Hence, we continue to proceed with precision and recall for top predicted neighborhoods.

Precision@L and Recall@L are shown in Figure 2.5. Model G-LRMM beats other components for both metrics greatly. For instance, G-LRMM model can achieve Precision@L of more than 0.75 in most cases. However, the Precision@L of alternative models is under 0.75 for the most part.

Visualization and Case Study The inferred probability is visualized as shown in Figure 2.6. Three different typical regions are also picked for case study research, as detailed in Table 2.2. Specifically, region A is an area around Zhongguancun (Chinese Silicon Valley), where many quality schools gather together. In contrast, most neighborhoods in region C are assigned to mediocre schools. Furthermore, the overall quality of region B lies between that of regions A and C. This area is actually composed of neighborhoods of different levels. The left inset of Figure 2.6 provides the comparison of price for three regions. For instance, region A has a higher median price than that of regions B and C. This confirms the underlying assumption for the impact of school districts on price to some extent. In a meanwhile, the subtle difference between regions B and C calls for further considerations other than price and necessitates the complexity of the proposed framework intuitively.

2.5 Discussion and Future Work

The proposed G-LRMM model is a general framework by integrating numerical price, textual comments and geographical location into the hierarchical probabilistic model. Since quality education resources might reduce the depreciation risk incurred

Table 2.2 School Districts for Regions A, B and C

Region	School district
A	Zhongguancun No. 1 (Good), Zhongguancun No. 2 (Good)
B	Zhanlan Road No. 1 (Good), Jinbu (Ordinary)
C	Sigenbai (Ordinary), Yutaoyuan (Ordinary)

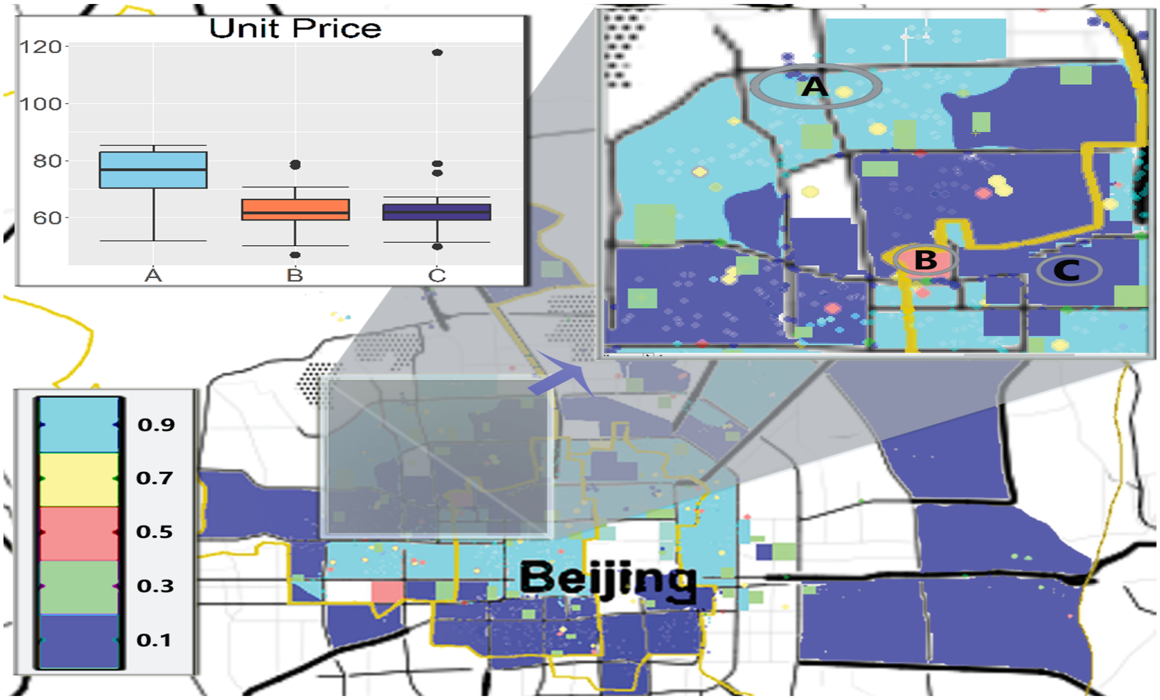


Figure 2.6 Distribution of school district quality.

by the real estate market's ups and downs, an alternative to deriving the hidden school district quality is based on the rate of return[8], which mainly characterizes the investment value of estate [55–57]. Thus, it remains our future focus to leverage the housing transaction records.

[8] https://en.wikipedia.org/wiki/Rate_of_return

Due to the limited accessibility of labeled data, our model is designed to infer latent variables, which can be regarded as an unsupervised learning scenario. If partial labeled data (training data) is available, the proposed model can also be revised to be a supervised version just by estimating parameters based on labels without extra complexity. Space constraints preclude a full discussion, but we also note that the supervised version outperforms classical supervised algorithms. The study of distance threshold and running time also justifies the robustness and feasibility of our model.

2.6 Conclusion

In this dissertation, we have developed real estate appraisal from the perspective of school districts for the first time. Then a geographical latent variable hierarchical probabilistic model is developed. The proposed G-LRMM model is able to capture heterogenous characteristics of neighborhoods. The comprehensive experiments are conducted on the real-world real estate market. The corresponding results show that our model can deliver the best performance over alternative methods with high feasibility. Besides, our work moves a step towards the modeling of heterogenous datasets.

CHAPTER 3

MODELING AND ELUCIDATION OF HOUSING PRICE

3.1 Introduction

It is widely acknowledged that the value of a house is the mixture of a large number of characteristics. House price prediction thus presents a unique set of challenges in practice. While a large body of works are dedicated to this task, their performance and applications have been limited by the shortage of long time span of transaction data, the absence of real-world settings and the insufficiency of housing features. To this end, a time-aware latent hierarchical model is developed to capture underlying spatiotemporal interactions behind the evolution of house prices. The hierarchical perspective obviates the need for historical transaction data of exactly same houses when temporal effects are considered. The proposed framework is examined on a large-scale dataset of the property transaction in Beijing. The whole experimental procedure strictly conforms to the real-world scenario. The empirical evaluation results demonstrate the outperformance of our approach over alternative competitive methods. We also group housing features into both external and internal clusters. The further experiment unveils that external component shapes house prices much more heavily than the internal one does. More interestingly, the inference of latent neighborhood value in our model is empirically shown to be able to lessen the dependence on the critical external cluster of features in house price prediction.

3.2 Related Work

This chapter [153] is an extended study of our preliminary results [152]. We thereby gave a brief introduction to the modeling of house prices and performed evaluation on one-period-ahead prediction accordingly [152]. this chapter entails four major extensions: (1) we perform long-span forecast research to further exhibit the feasibility and effectiveness of the proposed model; (2) we further perform hierarchical feature ablation analysis and elucidate dominant ingredients involved in the evolution of house prices, which has not been discovered yet; (3) we perform the further study on the way how the non-linear relationship among internal attributes shapes the predictive performance and the importance of influencing factors; (4) some case studies are presented to better the understanding of how the value of neighborhoods are associated with their desirability.

There have been other studies conducted on real estate appraisal. These works approached the problem from two perspectives roughly. One goes to the study of house ranking regarding the investment value of real estate [55–57, 172, 188]. The commonly used attributes are POIs [57,188], mobility behaviors [55,57,58], mixed land use [56] and community safety [172], among others. Interestingly, these features are closely associated with the desirability of the corresponding neighborhood [34]. They are reported to aid in real estate ranking remarkably. The other direction is dedicated to the modeling and prediction of property value itself [10, 27, 28, 65, 120, 144, 159].

Research in house price prediction falls into three categories roughly.

(1) The most well-known approach is hedonic pricing model based on regression analysis and its derivatives [66,118,159]. The methods of this kind simply incorporate

all available housing characteristics and estimate their contributions, which are relatively efficient in response to characteristics. The major limitation, however, is that a large number of hedonic attributes are needed to guarantee the performance. This is normally not the case in real-world situations. Thus, the performance is always constrained by the availability of informative features. Artificial neural networks (ANN) were also leveraged to model the complex nonlinearities among housing features [130]. Essentially, the hedonic regression model is functionally equivalent to a single layer ANN.

(2) The alternative strategy is to track market trends by utilizing homes sold multiple times without considering detailed housing features. The approaches of this sort are often called repeat sales methods. Since it was first proposed by [10], such a methodology has been further extended in different ways [27,28,65,144]. For example, the well-known S&P/Case-Shiller Home Price Index[1] is generated based on the arithmetic average of the repeat sales [144]. Regarding these methods, two prerequisites, however, have to be met. One is that no significant changes can be made to houses between sales. The other is that a large amount of data for each house is needed to fit the model. Here a short period can guarantee the first prerequisite but rule out the possibility of multiple sales of the same house and vice versa for the long period of time. These self-contradictory requirements definitely limit their applications. Most importantly, new houses without historical sales cannot be appropriately modeled in this manner.

[1] http://us.spindices.com/index-family/real-estate/sp-case-shiller

(3) Apart from purely repeat sales, some methods resort to neighboring houses and/or preceding sales by incorporating housing features simultaneously. Pioneering works including spatial autoregression [24, 34], temporal autoregression [120], and spatial-temporal autoregression [62, 127, 128, 147] were proposed. These studies have different kinds of limitations. Some economics-oriented works focused more on backward-looking index construction and parameter estimation without much emphasis on forward-looking predictability (see Refs. [24, 62, 127], and related references thereby). Some spatio-temporal lag algorithms [128, 147] were also proposed to model housing price trends. Particularly, the granularity of spatio-temporal correlation here is based on individual transactions of houses, where the overall transaction price of neighboring and precedent houses are directly leveraged for capturing the spatio-temporal correlation. An alternative modeling strategy is to regard the external component of a house price as a latent variable, which can be extracted by imposing spatial constraints on the corresponding "external" part of neighboring houses [34]. Based on repeat sales, an autoregressive model was proposed to construct home price index and then forecast the housing price [120], where random effects of local communities were introduced to infer the inherent land desirability implicitly. Random effects of this kind were also considered for estimating video-specific desirability in our previous study [154]. Different from the aforementioned methods, we model and track the land desirability of houses explicitly by imposing spatial-temporal constraints on the level of local community/neighborhood instead of individual houses. Regarding the experimental design, the scenarios of previous studies are not fully representative of the real-world situation. Simple synthetic

simulations based on artificial space-time process constructed on a square grid of housing sites with just a few hundreds of samples were performed or a subregions of a city with very limited samples and short transaction span were selected for study [147]. The most commonly empirical evaluation scenario adopted in existing spatiotemporal models is the random split of the whole dataset into training and test parts [34, 120]. The resultant issue here is that some future transaction records of houses are probably used to train the model and predict the market value of previously sold houses. In contrast to the aforementioned research, we strictly train the model based on historical transaction data and conduct the prediction in terms of one-period and multi-period-ahead situations. To address the issues of existing spatio-temporal models, we propose a new framework to leverage historical transaction data. Furthermore, the roles of different components and neighborhood value in the prediction of house prices are explored.

3.3 Price Modeling

3.3.1 Preliminary

The principal goal of our work is to model and predict house prices for the real-world situations. The value of a home is a real-valued variable, which is impacted by large numbers of characteristics [151]. Since house prices are dynamic and the real-world market value can only be observed when the transaction happens, the key point is the way historical transaction data are utilized. The overall framework is shown in

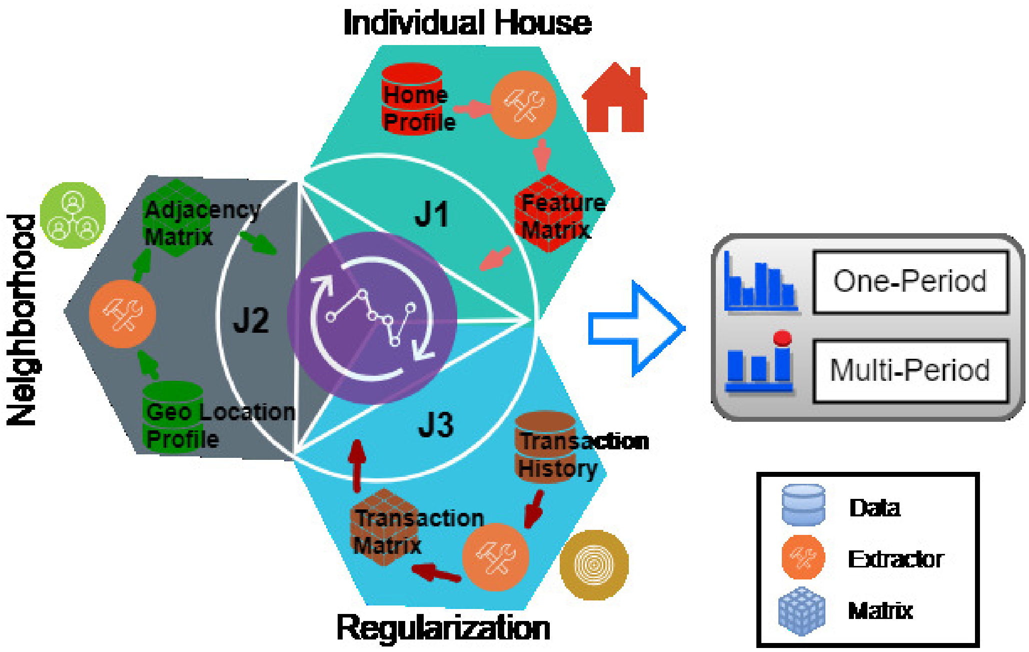


Figure 3.1 Framework overview of housing price prediction.

Figure 3.1. To further clarify this, we denote transaction matrix **R** as follows:

$$
\begin{array}{c|ccccc|cccc|}
i\backslash t & 1 & 2 & 3 & \cdots & T & T{+}1 & T{+}2 & \cdots & T{+}\Delta t \\
1 & 8 & 0 & 0 & \cdots & 6 & 0 & 9 & \cdots & 4 \\
2 & 0 & 7 & 2 & \cdots & 0 & 9 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
M & 0 & 0 & 4 & \cdots & 7 & 0 & 0 & \cdots & 5 \\
\hline
M{+}1 & 0 & 0 & 0 & \cdots & 0 & 8 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
M{+}\Delta i & 0 & 0 & 0 & \cdots & 0 & 0 & 3 & \cdots & 8
\end{array}
$$

where R$_{it}$ is the number of sold houses within neighborhood i at time t. The neighborhoods are ordered according to the time at which the first transaction happened. If time T is picked as the cutoff period, we then get the whole dataset split into four components. It is noted here that the bottom left component is composed of all 0s for those neighborhoods without transaction records prior to time T. Based on matrix R, we further derive one transaction indicator matrix **S** which is defined as follows:

$$S_{it} = \begin{cases} 1, & R_{it} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

In addition, another indicator matrix of missing value $\tilde{S}$ is denoted as

$$\tilde{S}_{it} = \begin{cases} 1, & R_{it} = 0, \sum_{i=1}^{t-1} R_{it} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

Obviously, $\tilde{S}_{it} = 1$ denotes that there are no houses sold at time period t, but at least one transaction happens prior to the current period (for example, R$_{12}$).

With those matrices, we formally define the problem of house price prediction as follows: given houses sold at each period ($1 \leq t \leq T$), we aim to predict the price of houses sold in the future. The common symbolic notation rule is obeyed throughout this article: upper case bold letters denote matrices, lower case bold letters indicate column vectors, and non-bold letters represent scalars.

3.3.2 Methodology

Model The house prices are always dynamic and change over time. The home's value is regarded as a combination of impacts of locality/neighborhood-induced attributes, and characteristics of an individual home. We introduce a time-varying latent variable, which is assumed to represent the impacts of all neighborhood-induced attributes on house value. The impact of time on the house itself is simply embedded as a regular attribute into our model. Thus, this task can be formulated as the following optimization problem. Mathematically speaking,

$$\arg\min_{\beta,U} \mathcal{J}_1 = \sum_{i=1}^{M}\sum_{t=1}^{T} S_{it} \sum_{k=1}^{R_{it}} (Y_{itk} - \beta X_{itk} - U_{it})^2 \tag{3.3}$$

where Y_{itk} is the value of home k in neighborhood i on time period t, $i \in \{1, 2, \ldots, M\}$, $t \in \{1, 2, \ldots, T\}$, $k \in \{1, 2, \ldots, R_{it}\}$. M is the number of neighborhoods, T is the number of sampled periods, and R_{it} is the number of houses sold within neighborhood i at time period t. $\boldsymbol{\beta}$ is a vector of coefficients for an array of all housing features. More discussions in regards to the functional format of those features will be presented in Section 3.5. In addition, U_{it} is the time-dependent desirability of neighborhood i at time period t, which is assumed to be highly related to the aggregating effects of external features in covariates $\mathbf{X}$ (we will validate this assumption in Section 3.4.3). The temporal evolution of $\mathbf{U}$ is thus the evolution of neighborhood value (desirability) across different transaction periods. $\mathbf{U}$ is actually the surrogate of the external component as mentioned in Section 3.1.

In real-world situations, the desirability of neighborhoods is generally constrained to the spatial-temporal interactions. To be specific, geographically close neighborhoods usually share similar location-associated characteristics, and thus their desirability is highly correlated. Meanwhile, the neighborhood value at time period t are closely related to that on the previous time period $t-1$. The smooth assumption of this kind is common in the temporal modeling [186]. To take into account both spatial dependency over geographical closeness and temporal dependency over temporal evolution, we design the following optimization procedure:

$$\begin{aligned}\arg\min_{\gamma_0,\gamma_1,U} \mathcal{J}_2 = &\sum_{i=1}^{M} S_{i1}(U_{i1} - B_{i1})^2 \\ &+ \sum_{i=1}^{M}\sum_{t=2}^{T} S_{it}(U_{it} - \gamma_0 B_{it} - \gamma_1 U_{it-1})^2\end{aligned} \tag{3.4}$$

where parameters γ_0 and γ_1 are the coefficients associated with neighbors' value B_{it} and its own prior value U_{it-1}, respectively. It is noted that when $t = 1$, we only have the spatial dependency. In particular, we have a time-dependent matrix $\mathbf{B} \subseteq \mathbb{R}^{M\times T}$, which is the aggregating value of adjacent neighborhoods. Thus, B_{it} is the corresponding value of neighbours of neighborhood i at time period t. To be specific, the weighted neighbors' value B_{it} is defined as follows:

$$B_{it} = \sum_{j=1}^{M} A_{ijt} U_{jt} \tag{3.5}$$

where $\mathbf{A} \subseteq \mathbb{R}_{\geq 0}^{M\times M\times T}$ denotes the interaction among different neighborhoods at different time periods. Thus, A_{ijt} is a weight to quantify the impact of neighborhood j on neighborhood i. The diagonal elements are specified as 0 such that a neighborhood itself is not involved. Regarding the weight A_{ijt}, we adopt the widely-used exponential kernel function based on the geodesic distance between neighborhood i and its neighbors j:

$$A_{ijt} = \frac{S_{jt}\text{exp}\{-qD^p(i,j)\}}{\sum_{l \in \mathcal{N}_t(i)} S_{lt}\text{exp}\{-qD^p(i,l)\}} \tag{3.6}$$

where p and q are nonnegative tunable hyper-parameters and $D(i,j)$ is the geographical distance between neighborhood i and j. The larger p and q are, the more important the distance plays the role in weights. If $p = 0$, the formula degrades to the arithmetic average. $\mathcal{N}_t(i)$ is the set of indices of K neighborhoods closest to i at time period t according to the geographical distance. Hyper-parameter K is introduced here to control the range of neighborhoods. Thus, A_{ijt} is specified to be 0 for those neighborhoods $j \notin \mathcal{N}_t(i)$.

To keep the surface smooth over both space and time and prevents the desirability from changing sharply, we also introduce an L_2 regularizer.

$$\mathcal{J}_3 = \sum_{i=1}^{M} \sum_{t=1}^{T} S_{it}(U_{it})^2 \tag{3.7}$$

Considering all the above analysis, we obtain our model based on the following optimization problem:

$$\begin{aligned}\underset{\gamma_0,\gamma_1,U,\beta}{\arg\min}\,\mathcal{J} &= \frac{1}{2}\sum_{i=1}^{M}\sum_{t=1}^{T} S_{it}\sum_{k=1}^{R_{it}}(Y_{itk} - \beta X_{itk} - U_{it})^2 \\ &+ \frac{\xi_0}{2}\sum_{i=1}^{M} S_{i1}(U_{i1} - B_{i1})^2 \\ &+ \frac{\xi_0}{2}\sum_{i=1}^{M}\sum_{t=2}^{T} S_{it}(U_{it} - \gamma_0 B_{it} - \gamma_1 U_{it-1})^2 \\ &+ \frac{\xi_1}{2}\sum_{i=1}^{M}\sum_{t=1}^{T} S_{it}(U_{it})^2\end{aligned} \tag{3.8}$$

where ξ_0 and ξ_1 are the regularization parameters. We can obtain latent matrix $\mathbf{U}$, coefficients $\boldsymbol{\beta}$ of housing features and spatial-temporal interaction coefficients γ_0 and γ_1 by solving Equation (3.8).

Learning Algorithm The inference of neighborhood value U_{it} for M neighborhoods across T time periods and parameter estimation must be carried out simultaneously. We propose the following iterative learning algorithm based on block-wise coordinate descent [20] to estimate latent variables and parameters. To be specific, the procedure of optimizing objective function $\mathcal{J}$ with respect to $\gamma_0, \gamma_1, \mathbf{U}, \boldsymbol{\beta}$ can be broken into two phases. In the first phase, we keep $\gamma_0, \gamma_1, \mathbf{U}$ fixed and minimize $\mathcal{J}$ with respect to $\boldsymbol{\beta}$. The second phase minimizes $\mathcal{J}$ with respect to $\gamma_0, \gamma_1, \mathbf{U}$ while keeping $\boldsymbol{\beta}$ fixed. The whole training procedure alternates between these two phases iteratively until convergence.

In regards to the first phase, as U_{it} is fixed, the estimation of coefficients β proceeds by minimizing $\mathcal{J}_1$ with respect to $\boldsymbol{\beta}$ as in Equation (3.3). $\mathcal{J}_1$ can be regarded as the least square error in regression analysis. Thus we leverage R software routine package 'lm' to estimate $\boldsymbol{\beta}$.

Regarding the second phase, we have the corresponding first derivatives of $\gamma_0, \gamma_1, \mathbf{U}$ as follows:

$$\frac{\partial \mathcal{J}}{\partial \gamma_0} = \sum_{i=1}^{M} \sum_{t=2}^{T} S_{it}(U_{it} - \gamma_0 B_{it} - \gamma_1 U_{it-1})(-B_{it}) \tag{3.9}$$

$$\frac{\partial \mathcal{J}}{\partial \gamma_1} = \sum_{i=1}^{M} \sum_{t=2}^{T} S_{it}(U_{it} - \gamma_0 B_{it} - \gamma_1 U_{it-1})(-U_{it-1}) \tag{3.10}$$

if $t = 1$

$$\begin{aligned} \frac{\partial J}{\partial U_{it}} &= -S_{it} \sum_{k=1}^{R_{it}} (Y_{itk} - \beta X_{itk} - U_{it}) \\ &- \xi_0 S_{it+1}(U_{it+1} - \gamma_0 B_{it+1} - \gamma_1 U_{it})\gamma_1 \\ &- \xi_0 \sum_{j=1}^{M} S_{jt} A_{jit}(U_{jt} - B_{jt}) + \xi_0 S_{it}(U_{it} - B_{it}) \end{aligned} \tag{3.11}$$

if $t > 1$

$$\begin{aligned}\frac{\partial J}{\partial U_{it}} &= -S_{it}\sum_{k=1}^{R_{it}}(Y_{itk} - \beta X_{itk} - U_{it}) \\ &- \mathbf{I}_t \xi_0 S_{it+1}(U_{it+1} - \gamma_0 B_{it+1} - \gamma_1 U_{it})\gamma_1 \\ &+ \xi_0 S_{it}(U_{it} - \gamma_0 B_{it} - \gamma_1 U_{it-1}) + \xi_1 S_{it} U_{it} \\ &- \xi_0 \sum_{j=1}^{M} S_{jt} A_{jit}(U_{jt} - \gamma_0 B_{jt} - \gamma_1 U_{jt-1})\end{aligned} \tag{3.12}$$

where $\mathbf{I}_t = 0$ if $t = T$, otherwise 1.

The update equations of γ_0 and γ_1 can be accordingly derived as

$$\gamma_0 \leftarrow \frac{\sum_{i=1}^{M}\sum_{t=2}^{T} S_{it}(U_{it} - \gamma_1 U_{it-1})B_{it}}{\sum_{i=1}^{M}\sum_{t=2}^{T} S_{it} B_{it}^2} \tag{3.13}$$

$$\gamma_1 \leftarrow \frac{\sum_{i=1}^{M}\sum_{t=2}^{T} S_{it}(U_{it} - \gamma_0 B_{it})U_{it-1}}{\sum_{i=1}^{M}\sum_{t=2}^{T} S_{it} U_{it-1}^2} \tag{3.14}$$

When $t = 1$, the update of U_{it} is given by

$$\begin{aligned}U_{it} \leftarrow &\frac{S_{it}\big[\sum_{k=1}^{R_{it}}(Y_{itk} - \beta X_{itk}) + \xi_0 B_{it}\big]}{S_{it}(R_{it} + \xi_0 + \xi_1) + S_{it+1}\xi_0\gamma_1^2 + \xi_0\sum_{j=1}^{M} S_{jt}A_{jit}^2} \\ + &\frac{S_{it+1}\xi_0\gamma_1(U_{it+1} - \gamma_0 B_{it+1}) + \xi_0\sum_{j=1}^{M} S_{jt}A_{jit}(U_{jt} - B_{jt}^{(-i)})}{S_{it}(R_{it} + \xi_0 + \xi_1) + S_{it+1}\xi_0\gamma_1^2 + \xi_0\sum_{j=1}^{M} S_{jt}A_{jit}^2}\end{aligned} \tag{3.15}$$

where $B_{jt}^{(-i)}$ is the weighted neighbors' value of neighborhood j with neighborhood i being excluded at time period t.

For $t > 1$, we have the following update equation

$$\begin{aligned} U_{it} \leftarrow & \frac{S_{it}\big[\sum_{k=1}^{R_{it}}(Y_{itk} - \beta X_{itk}) + \xi_0(\gamma_0 B_{it} + \gamma_1 U_{it-1})\big]}{S_{it}(R_{it} + \xi_0 + \xi_1) + \mathbf{I}_t S_{it+1}\xi_0\gamma_1^2 + \xi_0\sum_{j=1}^{M} S_{jt}\gamma_0^2 A_{jit}^2} \\ & + \frac{\mathbf{I}_t S_{it+1}\xi_0\gamma_1(U_{it+1} - \gamma_0 B_{it+1})}{S_{it}(R_{it} + \xi_0 + \xi_1) + \mathbf{I}_t S_{it+1}\xi_0\gamma_1^2 + \xi_0\sum_{j=1}^{M} S_{jt}\gamma_0^2 A_{jit}^2} \\ & + \frac{\xi_0\sum_{j=1}^{M} S_{jt}\gamma_0 A_{jit}(U_{jt} - \gamma_0 B_{jt}^{(-i)} - \gamma_1 U_{jt-1})}{S_{it}(R_{it} + \xi_0 + \xi_1) + \mathbf{I}_t S_{it+1}\xi_0\gamma_1^2 + \xi_0\sum_{j=1}^{M} S_{jt}\gamma_0^2 A_{jit}^2} \end{aligned} \tag{3.16}$$

where $\mathbf{I}_t = 0$ if $t = T$, otherwise 1. It is also important to update neighbor's value matrix $\mathbf{B}$ along with the latent variable matrix $\mathbf{U}$.

It is noted that the above two phases are only applicable for those neighborhoods with houses traded at specific time periods ($S_{it} = 1$). For those neighborhoods without records of traded houses ($\tilde{S}_{it} = 1$), we predict them based on the following strategy:

$$U_{it} = \gamma_0 B_{it} + \gamma_1 U_{it-1} \tag{3.17}$$

The main idea is to leverage the learned spatial-temporal integration coefficients to update them adaptively. Such updates can, in turn, impact the optimization procedure. In this manner, we can take full advantage of the spatial-temporal interaction. For those neighborhoods with $\tilde{S}_{it} = 0$ and $S_{it} = 0$, we ignore them as they cannot provide any useful information.

The overall procedure is summarized in Algorithm 2, and we call it TLHM for short. Also, we impose the nonlinearity on individual housing features. The revised algorithm is called TLHM_NL accordingly.

Algorithm 2: Time-aware Latent Hierarchical Model

```
Input  : distance matrix $\boldsymbol{D}$, covariate matrix $\boldsymbol{X}$, price vector $\boldsymbol{y}$
Output: latent variable matrix $\boldsymbol{U}$ and parameters $\gamma_0, \gamma_1, \boldsymbol{\beta}$
Initialize $U_{it} \leftarrow \frac{1}{R_{it}} \sum_{k=1}^{N_{it}} Y_{itk}$ for each pair $\{(i,t) : S_{it} = 1\}$;
Initialize $(\gamma_0, \gamma_1) \leftarrow (0.5, 0.5)$;
repeat
    $\beta \leftarrow$ Equation (3.3) ;
    $\gamma_0 \leftarrow$ Equation (3.13), $\gamma_1 \leftarrow$ Equation (3.14) ;
    for $\{$(i,t): $S_{it} = 1$ or $\tilde{S}_{it} = 1\}$ do
        if $S_{it} = 1$ then
            $U_{it} \leftarrow$ Eqs. (3.15) and (3.16);
        else
            $U_{it} \leftarrow$ Equation (3.17);
        end
    end
until convergence;
return updated $\boldsymbol{U}, \gamma_0, \gamma_1, \boldsymbol{\beta}$;
```

Prediction Inference With the preceding estimated parameters $\boldsymbol{\beta}$, γ_0, γ_1 and hidden time-dependent variables $\boldsymbol{U}$, which essentially uncovers both temporal and spatial interaction of the hidden neighborhood value, we obtain the learned model ready for prediction of a new house's market price sold in the future.

Suppose a house located in neighborhood i is traded at time period $t = T + \Delta t$ ($\Delta t \in \mathbb{Z}^+$), we have two scenarios to consider.

- $S_{iT} = 1$ or $\tilde{S}_{iT} = 1$ (top right component of matrix $\boldsymbol{R}$), the updated formula for neighborhood's value is defined as follows:

$$U_{it} = \gamma_0 B_{iT} + \gamma_1 U_{it-1} \tag{3.18}$$

- $S_{iT} = 0$ and $\tilde{S}_{iT} = 0$ (bottom right component of matrix $\boldsymbol{R}$)

$$U_{it} = B_{iT} \tag{3.19}$$

It is worthwhile to note that time period T of B_{iT} is fully representative of real-world situations for prediction, and we have no future transaction data of neighbors of neighborhoods. The overall prediction procedure is summarized in Algorithm 3.

3.4 Experiment and Elucidation

In this section, we describe our experimental procedure and report empirical evaluation results of the proposed algorithm on the real estate dataset of Beijing (DATASET AVAILABLE[2]).

[2] https://www.dropbox.com/s/isdw106x6hjwfkf/data_House_Price.csv?dl=0

Algorithm 3: One/Multi-period-ahead predictive Inference

```
Input  : matrix $\boldsymbol{U}$ and parameters $\gamma_0$, $\gamma_1$, $\boldsymbol{\beta}$, feature vector $\boldsymbol{x}$, time step
         $\Delta t$ between future period of interest and the latest training
         period
Output: house price $\hat{y}$
if $S_{iT} = 1$ or $\tilde{S}_{iT} = 1$ then
    for $t$ = $T$+1, $T$+2, ..., $T$+$\Delta t$ do
        $U_{it} = \gamma_0 B_{iT} + \gamma_1 U_{it-1}$;
    end
    $\hat{y} = U_{it} + \boldsymbol{\beta}^T \boldsymbol{x}$;
else
    $\hat{y} = B_{iT} + \boldsymbol{\beta}^T \boldsymbol{x}$;
end
return $\hat{y}$;
```

Table 3.1 Basic Statistics of Beijing Dataset

Items	Statistics
# of transactions	200,122
# of neighborhoods	5,487
# of administrative districts	12
time periods of transactions	01/2011 - 06/2015

Table 3.2 Statistics of Transactions and Their Neighborhoods across Different Quarters (Time Index Starts from $t = 1$ at 2011Q1 and Ends with $t = 18$ at 2015Q2)

Time period	2011Q1	2011Q2	2011Q3	2011Q4	2012Q1	2012Q2	2012Q3	2012Q4	2013Q1
# of Transactions	13	138	1,055	4,920	6,978	10,488	10,494	14,107	15,675
# of neighborhoods	13	126	767	2,002	2,357	2,783	2,887	3,217	3,318
Time period	2013Q2	2013Q3	2013Q4	2014Q1	2014Q2	2014Q3	2014Q4	2015Q1	2015Q2
# of Transactions	12,074	17,353	13,877	8,894	8,762	11,273	19,572	17,384	27,065
# of neighborhoods	2,986	3,469	3,310	2,848	2,745	3,004	3,605	3,585	4,145

3.4.1 Experimental Data and Preprocessing

We crawled historical real estate transaction data of Beijing. Few houses with extremely high unit price are excluded from our dataset. The basic statistics of our dataset are given in Table 3.1. Table 3.2 reports both number of transactions and neighborhoods over different time periods in details. For the same neighborhood, there are probably multiple transactions over different time periods. The preprocessing procedure of raw data is mainly elaborated in terms of house prices and detailed geographical location, respectively.

House Prices Since total house prices are the product of the footage square and the unit price for our dataset, the unit price is adopted as the prediction value of interest if not otherwise specified, which can better capture the underlying mechanism behind transaction data. Thus, terms "home's value", "house value", "unit price", "house prices" are interchangeable throughout this dissertation. House value ranges from tens of thousands CNY to hundreds of thousands CNY. Such a wide span results from a large variety of housing characteristics. As mentioned earlier, all of these features are categorized into two groups, namely, externality and internality, as described in Table 3.3. In the context of this dissertation, an administrative district[3] refers to a subdivision of a city, which is a government-controlled sub-city. For example, there are 12 administrative districts in the dataset of Beijing. Additionally, to ensure adequate data for each trading period, we divide entire time span into multiple three-month intervals or quarters. Different trading periods are assumed to indicate the market value of that period. It is worth noting that time effects might be confounded with age effects of a house. Thus an independently computed depreciation factor of age is also incorporated [26].

Geographical Location We convert the address of each neighborhood to the geographical coordinate (longitude and latitude) by use of Google Map Geocoding API[4]. Simultaneously, the cross-validation between the returned address and the neighborhood are conducted to guarantee the correctness of geographic data. With

[3] https://en.wikipedia.org/wiki/District_(China)

[4] https://developers.google.com/
maps/documentation/geocoding/intro

Table 3.3 The Housing Features and Groups

Group	Features
Externality	Administrative districts, Floor area ratio, Landscaping ratio, #Building, Subway access, Affiliated schools, Commercial environment
Internality	#Bedroom, #Living room, #House, Floor level, Orientation, Building type, Housing type, #Building floors, Age, Size

cleaned formatted geographical longitude/latitude coordination, we capitalize on Haversine formula[5], a formulation robust even for a small distance, to calculate the geodesic distance among neighborhoods. Finally, a distance matrix is obtained, and each entity represents the distance between any pair of neighborhoods.

3.4.2 Experimental Results

Evaluation Metrics In order to evaluate the performance of the proposed predictive model, we adopt widely used metrics for real-valued prediction problem [25, 34, 42, 82]. They are defined as follows:

- mean absolute value percentage error (MAPE) [25, 42],

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{3.20}$$

  where $\hat{y}_i$ is the predicted value and y_i is the ground-truth value.

[5] https://en.wikipedia.org/wiki/Haversine_formula

- median absolute value percentage error (MdAPE) [82],

$$\text{MdAPE} = \text{median}(\left|\frac{\hat{y_i} - y_i}{y_i}\right|), i \in 1, 2, \cdots, N \tag{3.21}$$

  MdAPE stays around where most of the data is and is thus more robust to outliers compared to MAPE.

- percentage of houses with absolute value percent error less than a predetermined threshold (PAPE@τ%) [34]. These metrics are exclusively introduced to measure performance with the cut-off value in contrast to the above two overall metrics. In this dissertation, the threshold τ is specified as 5 and 10 without loss of generality.

Baseline Algorithms In this section, we introduce alternative representative methods as baselines to justify the outperformance of the proposed model. (1) **LR**: The classical linear regression (hedonic pricing model) [159] incorporates all features listed in Table 3.3 plus transaction periods. The L2 penalty is placed on the objective function. The parameter estimation is coordinated by standard linear regression with elastic net penalty software routine *glmnet* in R language. (2) **ANN**: Artificial neural networks [130] share the same input features with **LR**. It is implemented with loss function of the mean squared error based on Keras[6]. (3) **RAA**: An arithmetic average of the repeat sales [144] was proposed to estimate home's value, which has been used to generate home price index as mentioned before. No adequate repeat sales of an individual home, however, are available given our dataset. Thus we adopt the revised

[6] https://keras.io

arithmetic average of similar houses within the same neighborhood as the estimator. If no historical sales with the neighborhood are provided, such an arithmetic average is performed on similar houses of the nearest neighborhood. (4) **KNN**: The procedure of predicting house prices involves two steps by considering both location information of neighborhoods and internal features of houses. We search for K_1 nearest ones from training neighborhoods (row 1 to M in matrix $\boldsymbol{R}$) based on geodesic distance. Then within those selected nearest neighborhoods, we collect K_2 most similar houses according to Euclidean distance in input space of houses' locality-oriented and internal features. The house price is the average over that of K_2 nearest houses. In the neighborhood search step, neighborhoods are confined to those which have traded houses during the latest time period (time period T in matrix $\boldsymbol{R}$). Such a setting ensures that locality, time period and internal features are considered as properly as possible under the framework of KNN. (4) **LME**: Chopra *et al.* proposed a static latent manifold estimation to capture unmeasurable desirability of neighborhoods [34]. In this dissertation, we incorporate the temporal effect as a discrete feature into LME for the fair comparison. For prediction, the effect of the latest trading time period is taken as future time effects. (6) **AR**: The autoregressive approach combines the fixed time effect, random location effects, and an autoregressive component [120]. (7) **STLAG**: The spatial-temporal lag model [127, 128, 147] considers spatial-temporal effects and impose autocorrelation effects on residuals over time.

One-period-ahead Prediction We pick houses sold in the final 2 years (2013Q3 $\sim$2015Q2) as benchmarks to evaluate the proposed method. Specifically, the experiments are set up as shown in Table 3.4. The comparison of various alternative

Table 3.4 Experiment Setting for One-period-ahead Prediction ($\Delta t = 1$)

Test Period	2013Q3	2013Q4	2014Q1	2014Q2	2014Q3	2014Q4	2015Q1	2015Q2
Training Period	2011Q1∼2013Q2	2011Q1∼2013Q3	2011Q1∼2013Q4	2011Q1∼2014Q1	2011Q1∼2014Q2	2011Q1∼2014Q3	2011Q4∼2014Q4	2011Q1∼2015Q1
M	4,638	4,804	4,915	4,991	5,047	5,104	5,184	5,272
T	10	11	12	13	14	15	16	17

Table 3.5 MAPE(MdAPE) and PAPE@5%(PAPE@10%) for One-period-ahead Prediction

MAPE(MdAPE)	**2013Q3**	**2013Q4**	**2014Q1**	**2014Q2**	**2014Q3**	**2014Q4**	**2015Q1**	**2015Q2**
LR	0.1796(0.1679)	0.1708(0.1534)	0.1684(0.1392)	0.1556(0.1232)	0.1672(0.1281)	0.1589(0.1248)	0.1645(0.1302)	0.1641(0.1336)
ANN	0.1417(0.1171)	0.1364(0.1044)	0.1481(0.1065)	0.1614(0.1231)	0.1550(0.1139)	0.1419(0.1124)	0.1478(0.1163)	0.1503(0.1180)
RAA	0.2265(0.2313)	0.2188(0.2208)	0.1995(0.1947)	0.1398(0.1277)	0.1058(0.0883)	0.1077(0.0927)	0.1089(0.0933)	0.1175(0.1037)
KNN	0.1117(0.0870)	0.1085(0.0800)	0.1181(0.0752)	0.1484(0.1101)	0.1280(0.0899)	0.1047(0.0781)	0.1023(0.0729)	0.1054(0.0780)
LME	0.0950(0.0772)	0.0880(0.0655)	0.0935(0.0619)	0.1305(0.1019)	0.1080(0.0866)	0.0801(0.0621)	0.0824(0.0630)	0.0856(0.0658)
AR	0.0954 (0.0774)	0.0879(0.0654)	0.0942(0.0623)	0.1304(0.1019)	0.1080(0.0868)	0.0804(0.0620)	0.0822(0.0632)	0.0863(0.0662)
STLAG	0.1206(0.0823)	0.1142(0.0714)	0.1140(0.06538)	0.1398(0.0996)	0.1255(0.0815)	0.0972(0.0647)	0.0975(0.0604)	0.1450(0.0644)
TLHM	0.0904(0.0699)	0.0860(0.0623)	0.0926(0.0583)	**0.1203**(0.0912)	0.1018(0.0741)	0.0807(0.0590)	0.0782(0.0589)	0.0849(0.0613)
TLHM_NL	**0.0878**(**0.0686**)	**0.0847**(**0.0611**)	**0.0880**(**0.0570**)	0.1820(**0.0827**)	**0.0917**(**0.0681**)	**0.0781**(**0.0588**)	**0.0748**(**0.0568**)	**0.0787**(**0.05970**)
PAPE@5%(PAPE@10%)	**2013Q3**	**2013Q4**	**2014Q1**	**2014Q2**	**2014Q3**	**2014Q4**	**2015Q1**	**2015Q2**
LR	0.1422(0.2950)	0.1527(0.3181)	0.1756(0.3575)	0.2164(0.4189)	0.2111(0.4066)	0.2088(0.4115)	0.1966(0.3944)	0.1960(0.3853)
ANN	0.2239(0.4309)	0.2480(0.4817)	0.2521(0.4740)	0.2268(0.4244)	0.2387(0.4482)	0.2296(0.4472)	0.2253(0.4382)	0.2268(0.4336)
RAA	0.0394(0.0976)	0.0461(0.1122)	0.07027(0.1636)	0.1875(0.3839)	0.2844(0.5572)	0.2741(0.5380)	0.2686(0.5329)	0.2316(0.4814)
KNN	0.2979(0.5653)	0.3234(0.6032)	0.3548(0.6194)	0.2405(0.4606)	0.2995(0.5432)	0.3370(0.6068)	0.3619(0.6328)	0.3383(0.6062)
LME	0.3312(0.6275)	0.3981(0.6874)	0.4157(0.7008)	0.2642(0.4940)	0.3002(0.5625)	0.4135(0.7160)	0.4115(0.7051)	0.3969(0.6846)
AR	0.3312 (0.6263)	0.4000(0.6885)	0.4169(0.6986)	0.2648(0.4935)	0.3012(0.5628)	0.4140(0.7164)	0.4129(0.7044)	0.3956(0.6818)
STLAG	0.3133(0.5891)	0.3583(0.6510)	0.3995(0.6752)	0.2808(0.5013)	0.3284(0.5812)	0.4040(0.6848)	0.4264(0.7160)	0.4021(0.6942)
TLHM	0.3660(0.6662)	0.4112(0.7126)	0.4394(0.7213)	0.2959(0.5371)	0.3540(0.6248)	**0.4386**(0.7227)	0.4330(0.7360)	0.4200(0.7132)
TLHM_NL	**0.3737**(**0.6804**)	**0.4176**(**0.7244**)	**0.4519**(**0.7381**)	**0.3244**(**0.5808**)	**0.3858**(**0.6634**)	0.4337(**0.7349**)	**0.4481**(**0.7544**)	**0.4310**(**0.7288**)

algorithms across eight trading periods is shown in Table 3.5. We have the following observations: First, the proposed method surpasses the alternative methods across almost all metrics overall. Concretely, TLHM model can predict around 43% of houses within an error margin of less than 5% and 74% of houses within an error margin of less than 10% on 2015Q1. This demonstrates the effectiveness of time-dependent latent value of neighborhoods. Second, nonlinear algorithms (e.g., ANN and TLHM_NL) achieve better performance than the linear counterparts (e.g., LR and TLHM) do as the former methods are capable of modeling the complex relationships among related features. Third, even though temporal effects have been modeled as a fixed feature into LME, TLHM still achieves performance gain over LME for all cases. To be specific, time effects are fixed for all latent desirability during the same period for LME, whereas they change across different neighborhoods. This indicates that simply combining temporal effects and location information is not enough for yielding an accurate prediction. Fourth, alternative spatio-temporal modeling algorithms STLAG and AR are inferior to our methods according to our empirical experiments. AR implicitly models the land desirability as a static random effects and constructs the temporal price index for individual housing features. STLAG is designed to capture the land desirability based on the overall price of neighboring individual houses. Lastly, both KNN and RAA beat the linear model remarkably for most cases even though the former methods use much fewer housing features than the latter. The commonality of KNN and RAA is to directly take advantage of location coordinates while the linear regression just leverages external location-associated features. This indicates that location information plays an important role in the predictive modeling of house prices as compared with internal features of individual houses. Thus, LR is more sensitive to housing features than the proposed model and KNN. The more in-depth comparison will be given in Subsection 3.4.3. For the nonlinear scenario, the comparison of this kind is also performed.

Multi-period-ahead Prediction We also present the comparison results of the long-span scenario, with which we predict the house prices for the next five quarters (i.e., $\Delta t = 5$). The experiment settings are detailed in Table 3.6. The prediction accuracy of different methods is reported in Table 3.7. The prediction performance

Table 3.6 Experiment Setting for Multi-period-ahead Prediction ($\Delta t = 5$)

Test Period	2014Q3	2014Q4	2015Q1	2015Q2
Training Period	2011Q1∼2013Q2	2011Q1∼2013Q3	2011Q1∼2013Q4	2011Q1∼2014Q1
M	4,638	4,804	4,915	4,991
T	10	11	12	13

of all methods degrades compared to one-period-ahead prediction in Table 3.5. As we use the predicted values from the past for future prediction, the problem of error accumulation is inevitable as the time gap increases. The main conclusions drawn from the one-period-ahead prediction still hold for multi-period-ahead cases.

It is noted that the performance gain is more salient in multi-period-ahead scenario. We attribute this to the capacity of our model of capturing the evolution of neighborhood value. This capacity becomes more obvious as the time span keeps increasing. Overall, the external component plays a more important role in shaping the housing price than the internal ones do as demonstrated in the following comparison experiments.

The above comprehensive evaluations jointly illustrate and justify the effectiveness of the proposed model.

Convergence Analysis and Parameter Tuning As shown in Figure 3.2, the convergence of the coordinate descent based iterative algorithm is very fast. To be specific, the proposed method tends to converge when the number of iteration approaches 30. The property of fast convergence has been reported in many algorithms related to coordinate descent [16, 42]. To tune hyper-parameters, we

Table 3.7 MAPE(MdAPE) and PAPE@5%(PAPE@10%) for Multi-period-ahead Prediction

MAPE(MdAPE)	**2014Q3**	**2014Q4**	**2015Q1**	**2015Q2**
LR	0.1574(0.1332)	0.1545(0.1262)	0.1636(0.1317)	0.1655(0.1351)
ANN	0.1491(0.1143)	0.1597(0.1229)	0.1783(0.1404)	0.1671(0.1287)
RAA	0.1736(0.1697)	0.1486(0.1407)	0.1365(0.1239)	0.1456(0.1319)
KNN	0.1218(0.0904)	0.1293(0.0981)	0.1552(0.121)	0.1529(0.1145)
LME	0.1026(0.0805)	0.1205(0.0970)	0.1421(0.1182)	0.1348(0.1071)
AR	0.1038(0.0815)	0.1213(0.0979)	0.1431(0.1200)	0.1337(0.1079)
STLAG	0.1440(0.0859)	0.1470(0.0918)	0.1659(0.1112)	0.2206(0.1016)
TLHM	**0.1008**(**0.0766**)	0.1184(0.0912)	**0.1299**(0.1052)	**0.1125**(0.0856)
TLHM-NL	0.1065(0.0786)	**0.1173**(**0.0904**)	0.1307(**0.1025**)	0.2237(**0.0841**)
PAPE@5%(PAPE@10%)	**2014Q3**	**2014Q4**	**2015Q1**	**2015Q2**
LR	0.2027(0.3903)	0.2082(0.4034)	0.195(0.3878)	0.1942(0.3781)
ANN	0.2327(0.4419)	0.2235(0.4235)	0.2002(0.3775)	0.2125(0.4008)
RAA	0.1012(0.2309)	0.1452(0.3277)	0.1857(0.3966)	0.1731(0.3662)
KNN	0.2973(0.5415)	0.2708(0.5086)	0.2162(0.4211)	0.2335(0.4454)
LME	0.3253(0.5984)	0.2679(0.5132)	0.2188(0.4302)	0.2420(0.4694)
AR	0.3204(0.5976)	0.2668(0.5098)	0.2158(0.4275)	0.2400(0.4696)
STLAG	0.3060(0.5683)	0.2971(0.5348)	0.2310(0.4567)	0.2644(0.4937)
TLHM	**0.3457(0.6178)**	0.2905(0.5389)	0.2437(0.4784)	0.3082(0.5675)
TLHM-NL	0.3370(0.6008)	**0.2927**(**0.5425**)	**0.2518**(**0.4887**)	**0.3100**(**0.5729**)

pick the last time period of training periods for validation. For example, for one-period-ahead prediction on 2014Q3, hyper-parameters are chosen based on the performance of 2014Q2. The optimal hyper-parameters are set as $p \in \{0,1\}$, $q \in \{1,3\}$, $K \in \{5,15\}$, $\xi_0 \in [0.5, 0.9]$ and $\xi_1 \in \{0.01, 0.1, 0.5\}$.

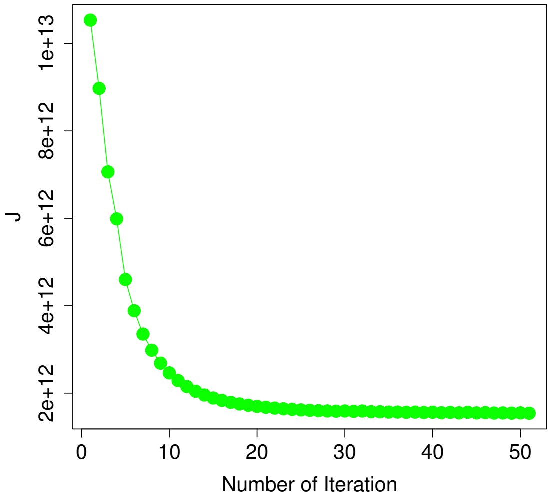


Figure 3.2 The evolution of objective value $\mathcal{J}$ over the number of iteration for training periods from 2011Q1 to 2014Q1.

3.4.3 Elucidation

Hierarchical Feature Ablation Analysis Apart from the proposal of a powerful predictive model, we also try to investigate roles of different components of housing features in price prediction. The basic procedure is to remove each group of features in turn from original models while preserving the remaining ones, then we compare the predication performance of revised models with the original one. The discrepancy among them can reflect the way those features impact house prices. Such an

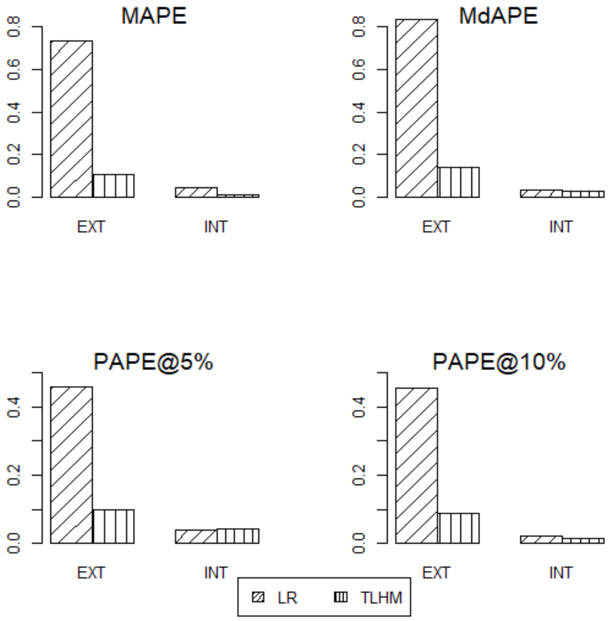


Figure 3.3 Performance comparison of different groups of features for LR and TLHM.

exploration can potentially provide useful insights into the collection and modeling of housing features. The linear regression is efficient in terms of the response to different characteristics of houses although it is not a good estimator for housing prices.

Thus, we mainly compare different groups of available features between the linear regression and the proposed TLHM model, as well as their corresponding nonlinear counterparts.

As shown in Table 3.3, we have two groups of features, i.e., externality and internality. To explore their impacts, we have the following settings: (1) **EXT**: the relative change of performance after excluding the external group of features from

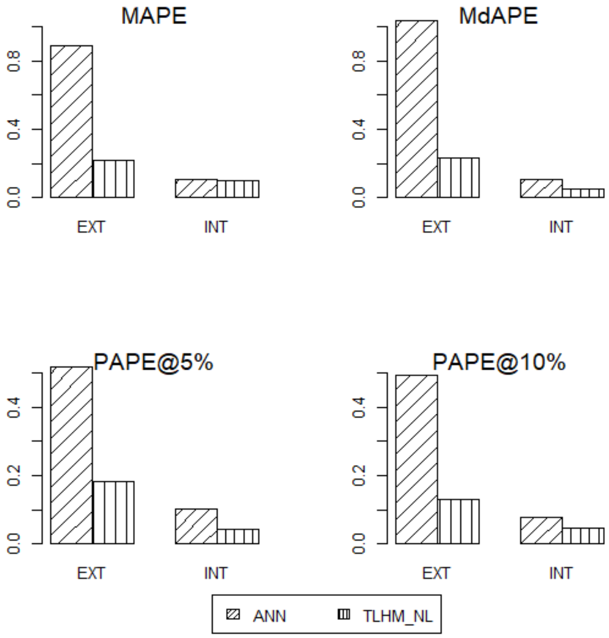


Figure 3.4 Performance comparison of different groups of features for ANN and TLHM-Nonlinear.

the original model; (2) **INT**: the relative change of performance after excluding the internal group of features from the original model.

We conduct the experiment with LR and TLHM as well as ANN and TLHM_NL for one-period-ahead prediction on 2014Q3.

The performance is comprehensively compared in terms of four different evaluation measures, as reported in Figure 3.3. For both LR and TLHM, it's found that different groups of features impact the prediction performance of house prices unequally. More specifically, the MAPE rises dramatically by around 70% after excluding an external group of features for LR. The impact of internal features, however, is marginal as compared to the exclusion of external features. Similar observations are also reported by more fine-grained metrics PAPE@%5, PAPE@%10. To some degree, it is revealed that external attributes are crucial components while the roles of internal home-specific components are relatively limited in this regard.

As with LR, external features shape the performance of TLHM model more heavily than internal ones do. The effect of external features on the proposed model, however, is degraded dramatically as compared to LR. Here, the inferred latent neighborhood value is able to capture external features partially since they are closely associated with the location. Furthermore, the impact of internal features are similar for both TLHM and LR. This might result from the limited role of the internal features. The way location information is utilized is also very important to the performance. The above analysis also holds true for the nonlinear scenario as indicated by the comparison between ANN and TLHM_NL in Figure 3.4.

Generally speaking, direct inference from geographical dependent information is better than external features. The reason might be that available external location-related features are always limited in real-world scenarios. This also partially shows that both methods RAA and KNN directly based on location information outperform LR and ANN with both internal and external location-related features when more locations are covered by training samples. Similar conclusions can be drawn from other time periods as well.

Altogether, the external component shapes the housing price more sharply than internal one does. Such dominant effects on house prices, however, can be replaced largely by the proposed surrogate neighborhood value under the framework of our model. Thus, it is indeed possible to infer a surrogate of the external component.

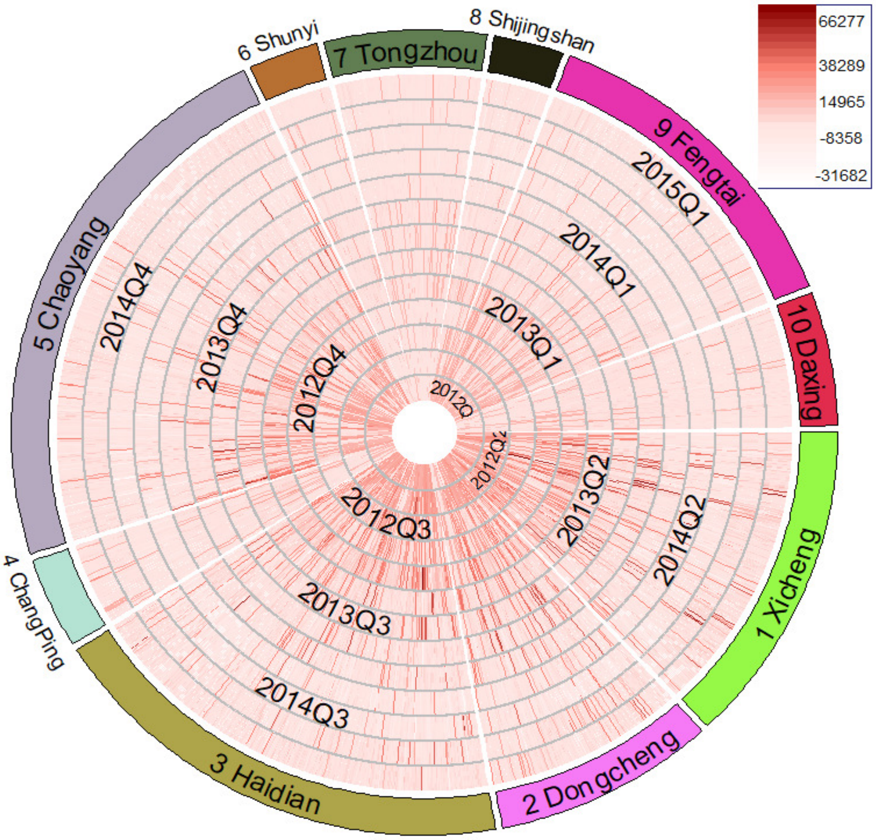


Figure 3.5 Circos plot of the evolution of the latent value of neighborhoods across time periods. The districts are Xicheng (1), Dongcheng (2), Haidian (3), Changping (4), Chaoyang (5), Shunyi (6), Tongzhou (7), Shijingshan (8), Fengtai (9) and Daxing (10).

Implication of Neighborhood Value Our method presents a unique perspective of house prices analytically by inferring time-varying neighborhood value. Figures 3.5 and 3.6 jointly illustrate the evolution of such inferred neighborhood value from transaction data, which could contribute to a better understanding of its implications conceptually. In Figure 3.5, we present the evolution of latent neighborhood value in ten main centric urban districts from 2012Q1 through 2015Q1 by using R circos plot package [69]. Each entity is the difference between latent values of two successive quarters. It can serve as a fluctuation index for house prices. Generally speaking, the latent value of neighborhoods goes up from 2012Q1 through 2013Q1, and then slows down from 2013Q2 to 2014Q1 which is followed by two declining quarters. Then the whole fluctuation ends up with a slight rise. Such an observation is consistent with real estate market trends in Beijing. To be specific, from December 2011 to July 2012, Beijing government issued a series of favorable loan policies. The price of houses rose faster than before accordingly. Afterwards, to suppress the property bubble, the government issued the "five policies and measures to regulate real estate market" on February 2013[7], which cooled down the acceleration of housing price. A series of tightened policies were operated continuously from May 2014 to October 2014 for regulating the property marketplace. The housing price kept going down during this regulating periods, which is then followed by the rise again after the end of the regulation[8]. It's noted that each neighborhood also has its own evolving pattern of neighborhood value. The evolving dynamics are actually beyond the representation of those fixed location associated features.

In Figure 3.6, the exact value of neighborhoods is geographically presented at time periods 2012Q1, 2013Q1, 2014Q1 and 2015Q1 via heat maps. As with Figure 3.5, it is shown that real estate market is basically going up. Interestingly, some hot spots are also found in the geographical heat maps. Here we pick areas A, B, and C for case studies, which seem to be desirable throughout the whole time periods of

[7] http://wiki.china.org.cn/wiki/index.php/five_policies_and_measures_to_regulate_real_estate_market
[8] http://www.sohu.com/a/131420084_651271

study. Area A is located in Zhongguancun (China's Silicon Valley), where the demand for houses is definitely large because this regional economic center offers many job opportunities. B is an area from College Road to Tsinghua High-Tech Park, where a couple of top-tier Chinese universities gather around. Furthermore, a lot of top school districts (elementary schools) in Beijing are located in areas A and B. Therefore, the population density is higher than surrounding regions. For area C, it stands out from suburban districts 4, 7, 8, 9 and 10. This results from the fact that the convenient transit facilities are available for residents to commute to the place of study/work in Beijing.

The above study demonstrates that the latent neighborhood value can capture location-oriented characteristics and reflects the prosperity of areas or even economic situation to some extent.

3.5 Discussion and Future Work

At the core of the proposed method is the inference of time-dependent latent neighborhood value at the granularity of the neighborhood. The prediction of future house prices is based on smoothness assumption on both spatial and temporal interactions. Admittedly, the real boundaries among different neighborhoods also shape the performance of model particularly for irregularly shaped regions [99, 137], which is beyond the scope of our work given our datasets. Furthermore, house prices are also impacted by the monetary policy, economic factors to some degree [4, 40]. As these factors are usually highly related to governmental regularization policies, they are less likely to follow the assumption of the spatiotemporal smoothness in our method. In this case, it is challenging to estimate their effects for future prediction just from historical transaction records of houses. These factors might jointly contribute to the difficulty of predicting house prices precisely. In turn, the proposed method can be readily adapted, and likely further improved as additional house features are provided. In addition, the potential connection among different neighborhoods can be explored under the framework of link prediction [104, 157].

The proposed framework is also very flexible. The regression component can be easily replaced by other types of methods (e.g., TLHM_NL in this dissertation) for potential performance improvement. Analogously, relative house ranking is often a concern for real estate investors. External characteristics are usually considered to promote the ranking performance such as POIs, mobility behaviors and so on [55]. Actually, POIs [13] and mobility behaviors [86] are closely related to the desirability of neighborhoods. So we raise an open question here: does the inferred desirability value of neighborhoods still work in other tasks of urban computing in terms of the replacement of external locality-associated features? This remains one avenue of future interest to explore.

3.6 Conclusion

In this chapter, we study the problem of house price prediction in real-world situations. A natural yet effective time-aware latent hierarchical model is proposed, where each neighborhood is associated with a set of latent variables that capture both spatial and temporal interactions among evolving house prices. The extensive experimental results demonstrate that latent hierarchical modeling of house prices with time-dependent effects can provide a powerful prediction capability as compared with alternative methods. It is also found that house prices are more sensitive to the external cluster of housing features than to internal ones. Furthermore, the proposed model can lessen the strong dependence on those crucial external characteristics by inferring the latent value of neighborhoods to some extent. The time-varying latent value is capable of capturing the desirability of neighborhoods on the price of individual houses and provides useful insights into local fluctuations of the real estate marketplace.

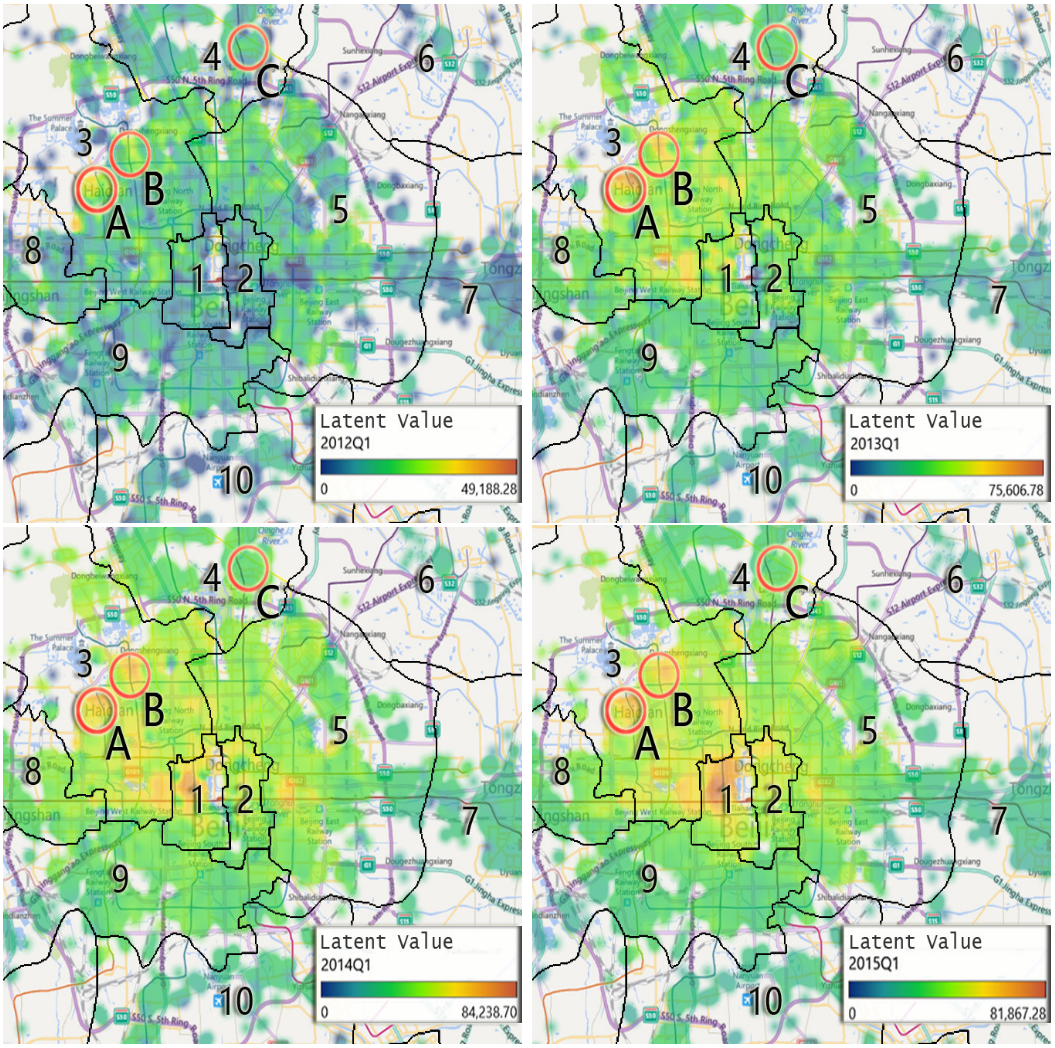


Figure 3.6 Geographic heat maps of latent neighborhood value at time periods 2012Q1, 2013Q1, 2014Q1, and 2015Q1. Area A is located in Zhongguancun (China's Silicon Valley), surrounded by top universities in China and high-tech companies. Area B is a region from College Road to Tsinghua High-Tech Park where a couple of top-tier universities in China gather around. Area C has the convenient transit facilities to downtown Beijing among residential neighborhoods in suburban districts.

CHAPTER 4

MODELING ITEM-SPECIFIC EFFECTS FOR VIDEO CLICK

4.1 Introduction

Prediction is widely employed to improve the number of video clicks and views, which are the key important indicators (KPIs) due to their contribution to revenue. The available predictive features, however, are generally limited as compared to the expected prediction capability from the algorithm side. Inspired by the intrinsic dependence among multiple clicks for the same video, we hypothesize that there exist some consistent effects involved in grouped click records. We then propose to recover such effects from the associated hidden features, which are likely to alleviate the insufficiency of features. The simulation studies are performed to elucidate how the derived grouped effects empower a model with additional discriminating capacity compared with the original one. The proposed methodology is further examined on the repository of PPTV (a leading video service provider in China) click records comprehensively. The results confirm the existence of the hypothesized effects and demonstrate their critical role in the performance improvement of video click prediction.

4.2 Related Work

Regression-based Scheme: There are a few approaches based on regression framework [129, 165]. In order to scale up the predictive algorithm, Vucetic *et al.* focused on a regression-related approach by considering relationships among items instead of similarities among users. Park *et al.* proposed a regression-based algorithm to avail of all information of users and items to construct pairwise feature space to predict item ratings for multiple cold-start cases. A regression-based latent factor model has also been developed to tackle the cold-start problem [3].

Item-specific Framework: Item-specific effects models have been widely employed to the longitudinal data analysis [175]. In contrast to such specific effects, the population-averaged effects are usually estimated by Generalized Estimating Equation (GEE) [71]. The latter integrates out individual effects to obtain average effects, which are equivalent to individual effects in linear case [98]. The basic idea of item-specific effects is also implicitly adopted by Koren's "BellKor's Pragmatic Chaos" final solution for predicting movie ratings. The solution won the Netflix grand prize [91] even though such effects were not explicitly claimed. However, our dissertation explores such effects under the regression framework analytically. To the best of our knowledge, this is the first attempt for video click prediction.

4.3 Methodology

4.3.1 Problem Statement

In this dissertation, our work is primarily focused on characterizing item-specific impacts involved in multiple click records of videos. If each record is supposed to be grouped into two categories, say, clicked and non-clicked, the problem can be cast as the integration of item-specific effects into the probability estimation: how likely the candidate videos get clicked.

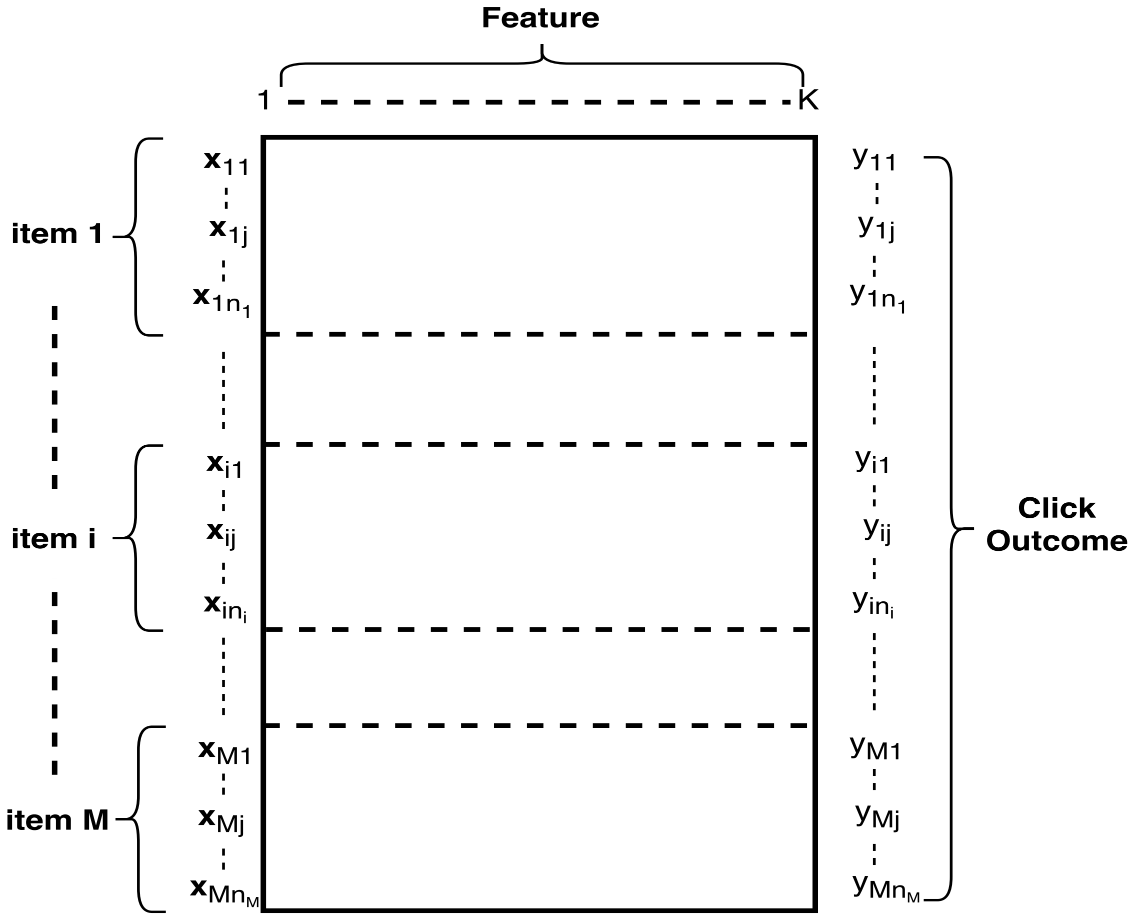
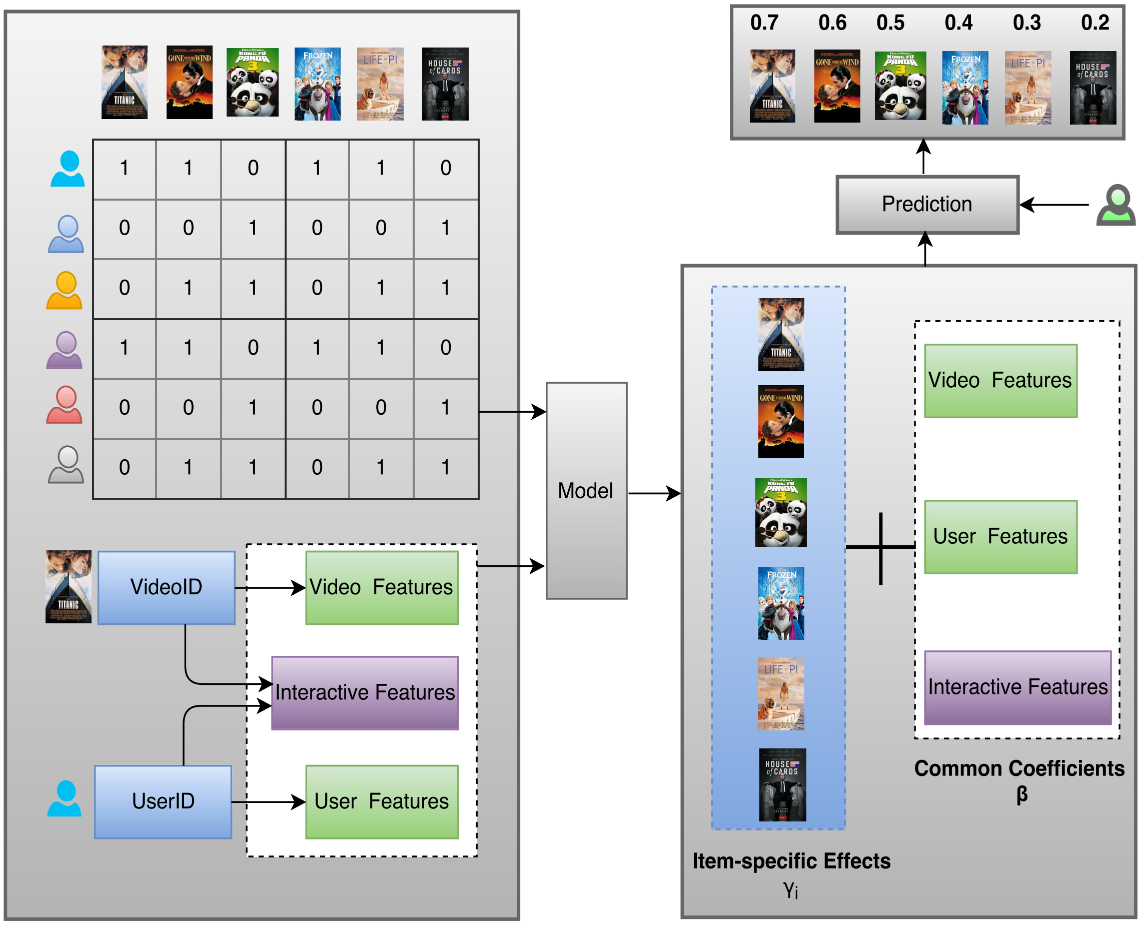


Figure 4.1 Schematic representation of variables.

4.3.2 Model

The logistic regression is commonly applied to the probability estimation of the binary response. However, the underlying assumption is that all observations are mutually independent, which also underpins many learning paradigms. This assumption does not hold true when different observations are associated with the same item (video in this dissertation) [85, 152]. Additionally, each video has some hidden attributes beyond the collected hand-crafted features. To take such dependence and limitation



1	1	0	1	1	0
0	0	1	0	0	1
0	1	1	0	1	1
1	1	0	1	1	0
0	0	1	0	0	1
0	1	1	0	1	1

Figure 4.2 The pipeline of the proposed method. Clicked and non-clicked recommended records are labeled as 1 and 0, respectively.

into consideration, we introduce item-specific effects into logistic regression. As the proposed method can be aware of such effects, we call it *item-specific effects aware* model or ISEA for short.

More formally, as shown in Figure 4.1, we have click outcome variable y_{ij}, a $K \times 1$ vector of features $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijk}, \ldots, x_{ijK})$, recorded at time $j = 1, \ldots, n_i$ for item $i = 1, \ldots, M$. Here M is the total number of videos and item i has n_i

historical click records by multiple users, there are thus a total of $N = \sum_{i=1}^{M} n_i$ click records. Feature vector $\boldsymbol{x}_{ij}$ includes video attributes, user features and interactive features between a user and a video as illustrated in Figure 4.2 and Table 4.3. We here further introduce γ_i to capture the item-specific effects.

Given γ_i, the two-state click variables $y_{i1}, \ldots, y_{in_i}$ are independent and follow a Bernoulli distribution

$$p(y_{ij}|\gamma_i) = \mu_{ij}^{y_{ij}}(1 - \mu_{ij})^{1-y_{ij}} \tag{4.1}$$

Obviously,

$$p(y_{ij} = 1|\gamma_i) = \mu_{ij} \tag{4.2}$$

In accordance with logistic regression and introduced item-specific effects, the *logit*[1] of underlying probability μ_{ij} is assumed to be a linear function of both collected features and item-specific terms

$$\text{logit}(\mu_{ij}) = \log(\frac{\mu_{ij}}{1 - \mu_{ij}}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \gamma_i \tag{4.3}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of common regression coefficients. Furthermore, under the framework of Generalized Linear Model [116,123], the conditional probability density function can be written as

$$f_y(y_{ij}|\gamma_i) = \exp\Big\{\frac{y_{ij} \cdot \xi_{ij} - b(\xi_{ij})}{a_{ij}(\phi)} + c_{ij}(y_{ij}, \phi)\Big\} \tag{4.4}$$

where $a_{ij}(\phi) = 1$, $c_{ij}(y_{ij}, \phi) = 0$, $\xi_{ij} = \text{logit}(\mu_{ij})$ and $b(\cdot) = \log(1 + \exp(\cdot))$.

[1] https://en.wikipedia.org/wiki/Logit

The model is intended to predict the click probability, thus, we can fit it based on likelihood analysis. Given item-specific effects γ_i, repeated click records from the same item are conditionally independent of one another. The probability density function of item-specific effects is represented as $f_\gamma(\cdot)$. Therefore, the joint probability density function of observed records for item i can be denoted as

$$f(\boldsymbol{y}_i, \gamma_i) = \prod_{j=1}^{n_i} f(y_{ij}, \gamma_i) = \prod_{j=1}^{n_i} f_y(y_{ij}|\gamma_i) f_\gamma(\gamma_i) \tag{4.5}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ is used to denote a $n_i \times 1$ vector of click outcome values of item i for brevity.

Essentially, item-specific effects γ_i serves as the impact of item i itself on associated repeated click records. It is noted that if γ_i is treated as the fixed unknown parameter, *Maximum Likelihood Estimation* might be inconsistent due to the fact that the number of unknown parameters increases with the number of items [124]. Furthermore, sampled items are thought to represent a population of items. Thus, item-specific effects γ_i is assumed to be drawn from a distribution. Because common coefficients $\boldsymbol{\beta}$ are estimated directly, item-specific effects are modeled as deviations from them, which have mean of zero. Thus, what is left to estimate is variance. In this chapter, effects γ_i is assumed to be a Gaussian random vector with mean 0 and variance σ^2, i.e., $\gamma_i \sim \mathcal{N}(0, \sigma^2)$ without loss of generality [117].

Since γ_i is latent, parameters $\boldsymbol{\beta}$, σ^2 are expected to be estimated by maximizing the integrated likelihood over γ_i as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{M} \int f(\boldsymbol{y}_i, \gamma_i; \boldsymbol{\beta}, \sigma^2) d\gamma_i \\ &= \prod_{i=1}^{M} \int f_y(\boldsymbol{y_i}|\gamma_i; \boldsymbol{\beta}) f_\gamma(\gamma_i; \sigma^2) d\gamma_i \\ &= \prod_{i=1}^{M} \int \prod_{j=1}^{n_i} f_y(y_{ij}|\gamma_i; \boldsymbol{\beta}) f_\gamma(\gamma_i; \sigma^2) d\gamma_i\end{aligned} \tag{4.6}$$

4.3.3 Parameter Estimation

The likelihood typically does not have a closed-form expression. In this case, the commonly used Laplace Approximation [9] is adopted to obtain an inexact form of likelihood. Then the derived likelihood is maximized to estimate the parameters.

The integral evolves as

$$\begin{aligned}&\int \prod_{j=1}^{n_i} f_y(y_{ij}|\gamma_i; \boldsymbol{\beta}) f_\gamma(\gamma_i; \sigma^2) d\gamma_i \\ &\propto \int \exp\Big\{ \sum_{j=1}^{n_i} \big[y_{ij} \cdot (\boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \gamma_i) - b(\boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \gamma_i) \big] - \frac{\gamma_i^2}{2\sigma^2} \Big\} d\gamma_i \\ &=: \int \exp\{Q(\boldsymbol{\gamma}_i)\} d\boldsymbol{\gamma}_i =: \mathcal{I}_i\end{aligned} \tag{4.7}$$

Suppose $Q(\gamma_i)$ has a global maximum at $\widehat{\gamma_i}$, then its first-order derivative vanishes at $\widehat{\gamma_i}$. In this case, $Q(\gamma_i)$ can be approximated to quadratic order Taylor aproximation

$$Q(\gamma_i) \approx Q(\widehat{\gamma_i}) + \frac{1}{2}(\gamma_i - \widehat{\gamma_i})^2 Q''(\widehat{\gamma_i}) \tag{4.8}$$

where $Q''(\widehat{\gamma_i})$ is given by

$$Q''(\widehat{\gamma_i}) = -\frac{1}{\sigma^2} - \sum_{j=1}^{n_i} b''(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \widehat{\gamma_i}) \tag{4.9}$$

Substitute Equation 4.8 into Equation 4.7 and the normal density integrates to 1, then we have

$$\begin{aligned}\mathcal{I}_i &\approx \exp\{Q(\widehat{\gamma_i})\} \int \exp\{\frac{1}{2}(\gamma_i - \widehat{\gamma_i})^2 Q''(\widehat{\gamma_i})\} d\gamma_i \\ &= \exp\{Q(\widehat{\gamma_i})\}(2\pi)^{\frac{1}{2}} \left|-Q''(\widehat{\gamma_i})\right|^{-\frac{1}{2}}\end{aligned} \tag{4.10}$$

The likelihood can thus be described by

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) \approx \prod_{i=1}^{M} \mathcal{I}_i = \prod_{i=1}^{M} \exp\{Q(\widehat{\gamma_i})\}(2\pi)^{\frac{1}{2}} \left|-Q''(\widehat{\gamma_i})\right|^{-\frac{1}{2}} \tag{4.11}$$

Item-specific effects $\widehat{\gamma_i}$ depends on the unknown parameters $\boldsymbol{\beta}$, σ^2. The numerical maximization of the likelihood will thus iterate between updating γ_i and ($\boldsymbol{\beta}$, σ^2). This procedure is coordinated by Bound Optimization by Quadratic Approximation (BOBYQA) [131], which seeks the least value of a nonlinear function subject to bound constraints, without using derivatives of the likelihood. The procedure is detailed in Algorithm 4.

4.3.4 Click Inference

With the foregoing estimated parameters $\boldsymbol{\beta}$, σ^2 and hidden variable γ_i, which essentially capture both effects of available features and item-specific effects, we get the learned model ready for click prediction of a new user-video interaction. To be

```
Algorithm 4: ISEA
  Input  : feature vector $\boldsymbol{x}_{ij}$, click outcome $y_{ij}$, $i = 1, \ldots, M, j = 1, \ldots, n_i$
  Output: parameters $\boldsymbol{\beta}$ and $\sigma^2$, and item-specific effects $\gamma_i$, $i = 1, \ldots, M$
  initialize $\boldsymbol{\beta}$ by logistic regression;
  initialize $\sigma^2 \leftarrow 1$;
  initialize $\gamma_i$ by drawing from $\mathcal{N}(0, \sigma^2)$ ;
  $\mathcal{L}(\boldsymbol{\beta}, \sigma^2) \leftarrow$ Equation 4.11;
  update $\boldsymbol{\beta}$, $\sigma^2$ and $\gamma_i$ by maximizing $\mathcal{L}(\boldsymbol{\beta}, \sigma^2)$ with BOBYQA;
  return $\boldsymbol{\beta}$, $\sigma^2$, $\gamma_i$;
```

specific, the probability of video i clicked at times j is generated by aggregating effects $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ and item-specific effects γ_i, where $\boldsymbol{x}_{ij}$ is a vector of collected features (for example, features listed in Table 4.3).

4.4 Experiment

In this section, we first perform a series of simulation studies to elucidate the item-specific effects of this kind on click prediction. The offline evaluation procedure is then described and the empirical results of the proposed methodology on the PPTV video click dataset are reported accordingly.

4.4.1 Simulation Studies

We first present simulation results to demonstrate the proposed methodology. The synthetic data are generated on the basis of the following simple formula:

$$\mathrm{p}(y_{ij} = 1) = \frac{1}{1 + e^{-\beta_0 - \boldsymbol{x}_{\boldsymbol{ij}}^{\boldsymbol{T}} \boldsymbol{\beta}}} \tag{4.12}$$

where $y_{ij} = 1$ if $p(y_{ij}) \geq p^*$, $y_{ij} = 0$ otherwise. Here p^* is the probability threshold.

To facilitate the understanding of simulation studies, we set n_i as a constant of N as mentioned in Section 4.3.2. We thus have a total of $L = M \times N$ samples here. Put another way, we can image that M videos with N clicking labels per one are recorded. We generate $K = 100,000$ different clicking samples with $M = 1,000$ and $N = 100$ independently where each predictive feature x_{ijk} is drawn from a Gaussian distribution. To simulate the intrinsic item-specific effects involved in clicking behaviors, we further impose such effects on the first two predictive features x_{ij1} and x_{ij2}. Then M item-specific effects are also drawn independently from a Gaussian distribution for both features, respectively. Mathematically speaking, we proceed with the following procedure:

1. For each clicking feature k, draw $x_{ijk} \sim \mathcal{N}(0, 1)$

2. For item-specific effects η_{i1} and η_{i2},

   (a) $\eta_{i1} \sim \mathcal{N}(0, \rho)$

   (b) $\eta_{i2} \sim \mathcal{N}(0, \rho)$

3. For clicking features x_{ij1} and x_{ij2},

(a) $x_{ij1} = x_{ij1} + \eta_{i1}$

(b) $x_{ij2} = x_{ij2} + \eta_{i2}$

where ρ is referred to as the standard deviation. In this case, multiple click samples share same effects of this kind via the introduced η_{i1} and η_{i2} if they are recorded from the same video i.

Without loss of generality, the synthetic samples are assumed to be composed of $K = 5$ features. Since the underlying Gaussian distributions are all with mean of 0, we have the associated coefficients $\boldsymbol{\beta}$ be a vector of (4, -1, -3, -0.5, -2), and intercept $\beta_0 = 2.5$ such that their summation is equal to 0. Also, the probability threshold is set to be $p^* = 0.5$. In this case, it is expected to roughly generate balanced positive and negative samples for this study. It is noted that ρ controls the strength of item-specific effects. The higher ρ is, the stronger the resulting effects are. We generate data with different $\rho = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ under the constraint of being less than or equal to 1. It avoids the possible exaggerated role of such effects via a setting akin to what can be found in practice.

Table 4.1 Experiment Settings

Experiment	E1	E2	E3	E4
Hidden Features	x_{ij1}	x_{ij2}	x_{ij1}, x_{ij2}	x_{ij3}

We perform four different simulation experiments over varied standard deviation ρ. We simulate different hidden features by excluding them alternately as detailed in Table 4.1. Take E_1 with x_{ij1} being a hidden feature for example, we exclude x_{ij1} and leverage the remaining 4 features to perform model estimation and inference.

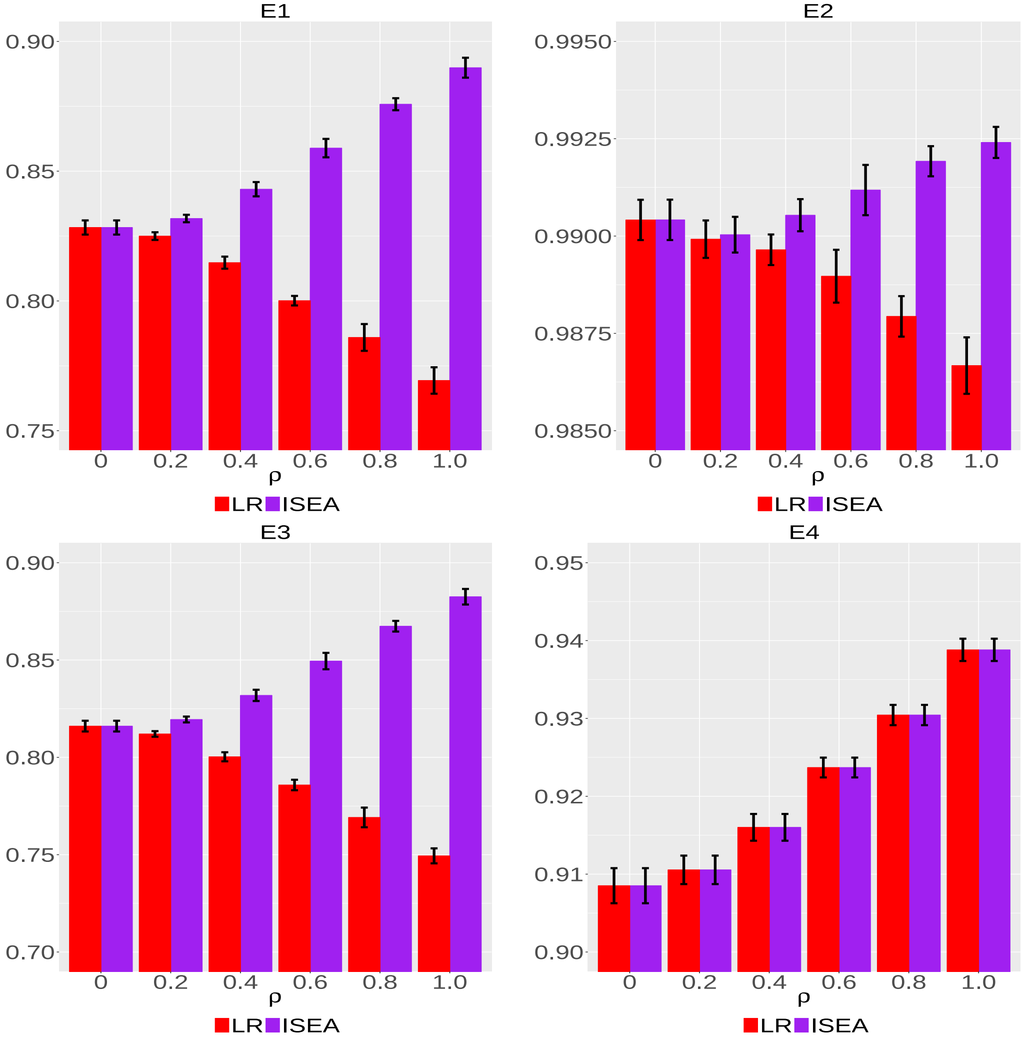


Figure 4.3 AUC@ROC of ISEA and LR across varied standard deviation ρ under E_1, E_2, E_3 and E_4. The height is mean of 10 independent runs and the error bar is the standard deviation of 10 replicates.

All simulations are repeated 10 times to assess the classical Area under ROC curve (AUC@ROC). As a comparison, we also perform the analysis of standard logistic regression (LR) on the simulated data, which doesn't account for item-specific effects. We present the comparison results over 10 replications as shown in Figure 4.3. For

E1, $\mathbf{E_2}$ and $\mathbf{E_3}$, the proposed ISEA outperforms LR greatly. To be concrete, the stronger the effects (standard deviation ρ) are, the higher the superiority of our method is. Specifically, when $\rho = 0$, both models achieve exactly same results for different experiments. In addition, as ρ progressively increases, item-specific effects get increasingly large weights into the click status. Thus, ISEA becomes better along with the capacity of recovering item-specific effects, whereas LR deteriorates due to more weights are excluded. What's more, the disparities among different experiments for same models indicate that the deterioration of model predictive performance is also associated with the increasing importance of hidden features. Regarding $\mathbf{E_4}$, we observe identical predictive performance of both methods. This is natural as x_{ij3} involves no grouped effects. Therefore, the proposed methodology is able to improve predictive performance provided that hidden features are indeed associated with item-specific effects.

4.4.2 Real Data and Preprocessing

To prepare for the real-world evaluation, we generate the dataset with ground-truth labels from PPTV video click repository as described in the following procedure. When a user clicks one video in the playlist, the user and all videos including non-clicked ones in the same playlist will be recorded along with features listed in Table 4.3.

Table 4.2 Basic Statistics of PPTV Video Click Records

# of records	# of videos	# of users	#negative	#positive
69,284	3,718	55,806	52,989	16,295

Table 4.3 Feature Profile of PPTV Video Click Records

Category	Feature	Format	Remarks
User	Type	Categorical	VIP, registered, unregistered
	Duration	Numerical	total duration of clicked videos before 10PM (unit: second)
	#Watch	Numerical	total number of clicked videos before 10PM
	#Search	Numerical	total number of searches before 10PM
	#Click	Numerical	total number of clicks before 10PM
Video	Group	Categorical	26 categories: news, movie, animation, fashion, etc.
User-Video Interactive	Device	Categorical	platform: iPhone, iPad, Android Phone
	TimeSlot	Categorical	watching time slot: 10PM-11PM, 11PM-12PM
	CtgDrt	Numerical	duration of videos at level of User-Device-Category-TimeSlot
	#CtgClk	Numerical	number of clicks at level of User-Device-Category-TimeSlot
	#Search	Numerical	number of searches at level of User-Device-TimeSlot
	#RecClk	Numerical	click number of videos at level of User-Device-TimeSlot
	#Comment	Numerical	number of comments at level of User-Device-TimeSlot
	#BulletScr	Numerical	number of bullet screen at level of User-Device-TimeSlot

The combination of the clicked video, the user and the associated features is labeled as a positive sample. To reduce position-induced bias, we pick those non-clicked videos located at the top of the playlist as negative samples. After filtering out duplicated records, we secure the refined samples. The basic statistics are listed in Table 4.2. It is noted that each video is associated with about 18.6 users and each user has only around 1.25 click records on average. As shown in Table 4.3, all features of the dataset are grouped into three categories, say, user profile, video profile, user-video interactive features. For videos, only coarse-grained categories are available.

4.4.3 Experimental Results

The whole sampled click records are randomly shuffled and split into training, validation and test datasets with ratio of 8:1:1. Due to the imbalance between clicked positive records and non-clicked negative records, the splitting procedure is conducted on positive and negative records separately. Then two kinds of records are concatenated to form the stratified training, validation and test datasets. The proposed model and baselines are all learned on training data. The validation data are utilized to coordinate the fine-tuning of hyper parameters if applicable. The corresponding evaluation metrics, baseline algorithms and the performance analysis are detailed in the following subsections respectively.

Evaluation Metrics In order to evaluate the prediction performance of the proposed methodology, we adopt AUC@ROC, AUC@PR (Precision-Recall curve) and probability cutoff based recall, precision, F1 score, accuracy, Matthews correlation coefficient (MCC) [113]. Two alternative metrics are also introduced, which focus on the top predicted records instead of the overall records. Specifically, Precision@L and Recall@L for top L predicted records are frequently used [140, 151]. Formally, given top L predicted records $\mathrm{R_L}$ sorted in the descending order of the predicted scores, they are defined as

$$\begin{aligned} \mathrm{Precision@L} &= \frac{|\mathrm{R_L} \cap \mathrm{R_C}|}{\mathrm{L}} \\ \mathrm{Recall@L} &= \frac{|\mathrm{R_L} \cap \mathrm{R_C}|}{|\mathrm{R_C}|} \end{aligned} \tag{4.13}$$

where $\mathrm{R_C}$ is a set of clicked positive records. Thus, Precision@L is actually an offline approximation to overall click rate.

Baseline Algorithms In this section, we will introduce other algorithms as baseline schemes to demonstrate the appealing role of item-specific effects. (1) **LR**: A predictive logistic regression model considers features listed in Table 4.3. (2) **PLR**: A pairwise logistic regression is fitted by constructing joint feature space for user/video pairs in addition to the features adopted by method LR. The regression coefficients are optimized with elastic net regularization. Two hyper parameters have $\alpha = 0.9$ and $\lambda = 0.00013$, respectively [129]. (3) **GBM**: Gradient boosting machine is proven to be a very successful framework of gradient boosted decision trees for solving a real-world ranking task [21]. The number of trees and learning rate are 389 and 0.043, respectively. All of these hyper parameters are obtained by grid search on validation dataset. The focus of the real-world experiments is to demonstrate the existence of item-specific effects in video click, which can boost the performance of regression methods. We don't aim to beat other powerful algorithms; however, this chapter can be extended to improve them in a proper way.

Table 4.4 Cost Matrix

	actual negative	actual positive
predicted negative	c_{00}	c_{01}
predicted positive	c_{10}	c_{11}

Overall Performance and Analysis Since evaluation metrics: recall, precision, f1 score, accuracy, mcc are based on specified cutoff probability, we have the following threshold with Bayes minimum risk [48]. The cost matrix is given in Table 4.4 for reference. Generally speaking, in an unbalanced dataset, false negative (missing a potentially clicked video) is more costive than false positive (misclassifying a non-clicked video as clicked one). For the correct prediction, the corresponding cost is zero naturally, $c_{00} = c_{11} = 0$. As we have no information about cost of misclassification from the commercial side, we adopt a general strategy and let $c_{01} = p(-)$, $c_{10} = p(+)$, where $p(-)$ and $p(+)$ are priors of negative and positive samples in training dataset, respectively [38]. Within the framework of Bayes minimum risk, the optimal threshold is given as follows:

$$p^* = \frac{c_{10} - c_{00}}{c_{10} - c_{00} + c_{01} - c_{11}} = p(+) \tag{4.14}$$

Since training and test datasets are drawn from the same population with stratified sampling in our setting, the optimal threshold in test dataset holds same, which is

$$p^* = \frac{\#(\text{positive samples})}{\#(\text{negative samples}) + \#(\text{positive samples})} \tag{4.15}$$

The comparison results amongst the proposed model and the baseline schemes are examined in Table 4.5. Overall, we observe that our method significantly dominates other baseline algorithms. First of all, even though our method achieves no superiority in terms of recall, it still has a commanding lead regarding precision. The precision of our method is more than 0.44, which is much better than both regression based algorithms and GBM. The comprehensive metric F1 score further supports the superiority of our method against counterpart ones. Another observation is that the outperformance of the proposed method still holds true for accuracy and MCC. Accuracy is sensitive to imbalance between positive and negative samples and is likely to be misleading to some extent. In our dataset, the positive clicked records only constitute less than 25% of the whole test dataset. As opposed to accuracy, MCC is generally regarded as balanced measure, which makes our results more convincing.

As shown in Figure 4.4, we also generate receiver operating characteristic (ROC) curves discriminating clicked records against non-clicked ones on the test data. ISEA performs significantly better than alternative methods, leading to more than 15% improvement in the AUC@ROC relative to GBM. This suggests the importance of accounting for item-specific effects involved in videos. An alternative metric to evaluate the performance is Area under Precision-Recall Curve (AUC@PR) as given by Figure 4.5 [39, 151]. Actually, the number of non-clicked samples dwarfs the number of clicked ones here. AUC@PR is more sensitive to reflect the detection of clicked samples than AUC@ROC is because neither precision nor recall considers true negative samples [39]. As expected, the absolute improvement is over 20% and the relative improvement is over 50% under the AUC@PR metric. Thus, it is reasonable to say that the proposed scheme indeed enjoys the dramatic advance of click prediction.

Aside from the overall metric on dataset, the recall and precision for top predicted records also play an important role in evaluation of click prediction. As shown in Figure 4.6, the proposed model beats alternative methods by a significant margin. It is capable of achieving precision of more than 0.6 at the minima. By contrast, the Precision@L of other three models absent from item-specific effects is less than 0.5 for most cases. Moreover, regular regression-based methods achieve pretty similar performance here. It is hard to improve them by penalty and gradient boosting. Finally, a decreasing trend of precision for the proposed model can be seen as L keeps increasing. Usually for a fixed candidate list, negative records are much more than positive ones. For a reasonable algorithm, the latter ones are more likely to be ranked at top positions in terms of probabilities. Therefore, precision out of a

Table 4.5 Performance for ISEA and Baselines

Model	Recall	Precision	F1 Score	Accuracy	MCC
LR	**0.72867**	0.33296	0.45707	0.59296	0.23751
PLR	0.72437	0.33249	0.45577	0.59324	0.23527
GBM	0.68140	0.35294	0.46502	0.63135	0.25329
ISEA	0.72560	**0.44453**	**0.55131**	**0.72229**	**0.38968**

set of L samples (as defined in formula 4.13) is expected to decrease along with the increment of set size L. As with Precision@L, similar dominance of method ISEA can be observed on Recall@L as in Figure 4.7.

In summary, the above results along with simulation studies show that there indeed exist some grouped effects in video click. The derived item-specific effects can also be interpreted as a surrogate or recovery of uncollected profiles of videos or video-user interactive characteristics to some extent.

4.5 Discussion and Future Work

In this dissertation, we propose to infer item-specific effects beyond limited features, which would offer a unique perspective of video click prediction. Our work also alleviates the problem caused by insufficient features and sparse interactions between users and videos by deriving such effects, which is demonstrated and confirmed by empirical studies clearly. For click prediction and ranking, there have been a number of algorithms designed for different circumstances. For example, BPR-based method is powerful given a reasonable user-item interaction matrix (for example, each user

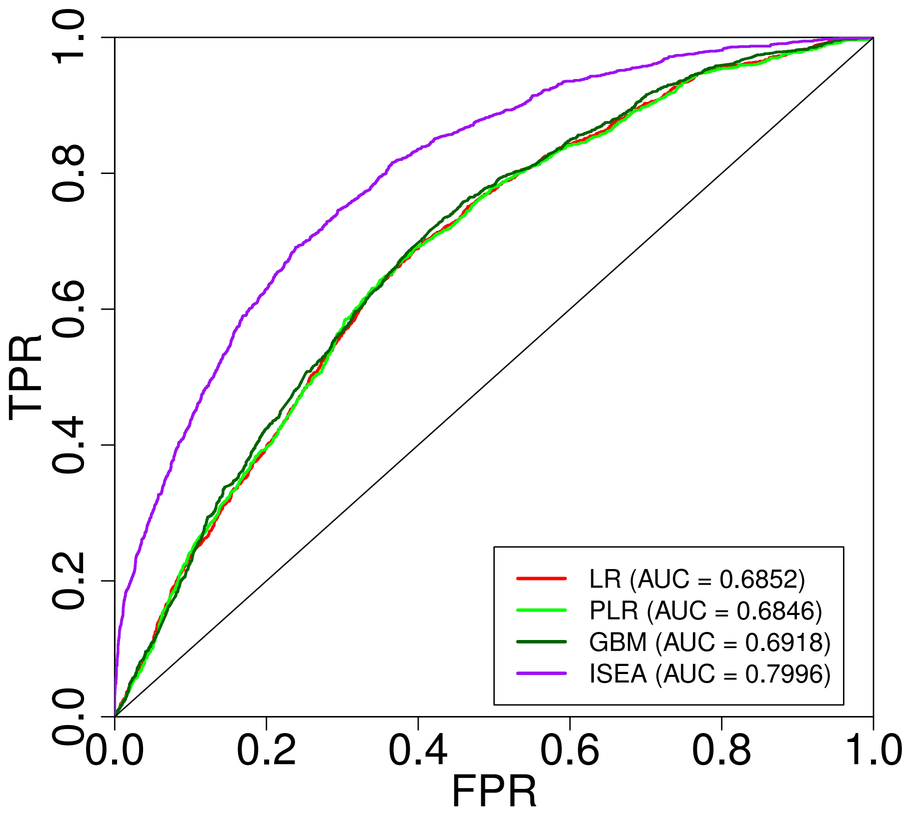


Figure 4.4 ROC curves comparison.

has at least 10 items, and each item has at least 10 users as stated in [132]). Our model still outperforms BPR-based method by a significant margin for the dataset used in this dissertation. That being said, the main goal of our work is to emphasize the important role of item-specific effects in click prediction instead of beating other schemes proposed for different scenarios. However, our research also raises an open question: how to improve the existing algorithms by considering such effects? This might be a promising research direction.

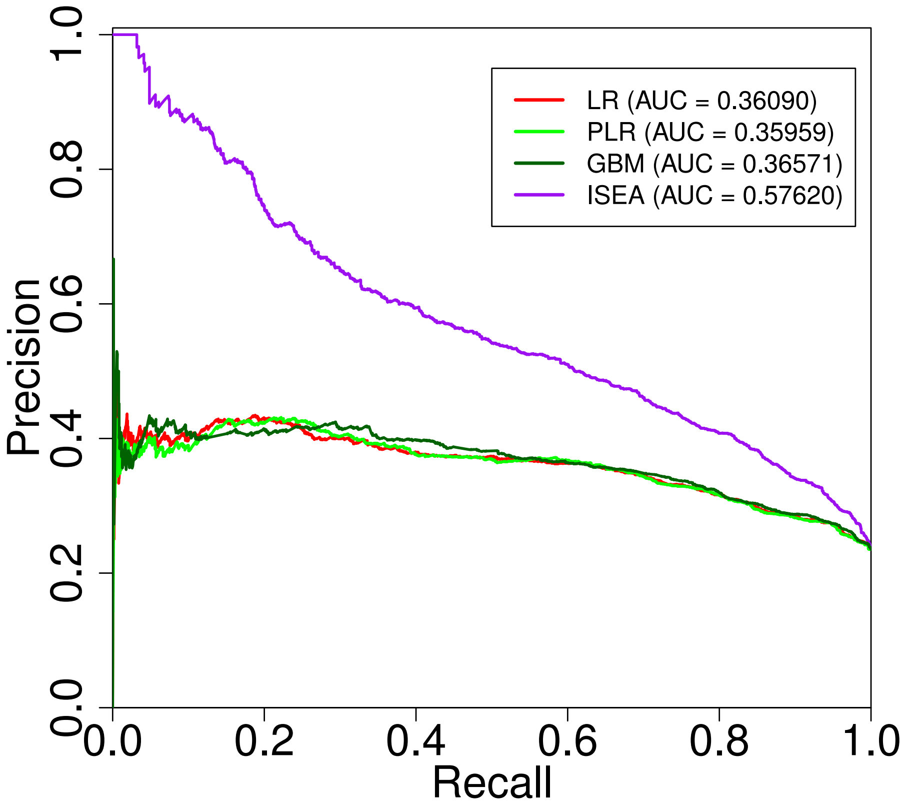


Figure 4.5 PR curves comparison.

As mentioned earlier, unlike the direct estimation of regression coefficients, item-specific effects are assumed to be sampled from a normal distribution. Such an assumption is reasonable since the effects are much less sensitive to distributional assumptions. In other words, the effects are still close to the truth even though the true distribution of the random effects is non-normal[2].

[2] http://www.statistik.lmu.de/institut/ag/biostat/teaching/lots2007/GLME2-4.pdf

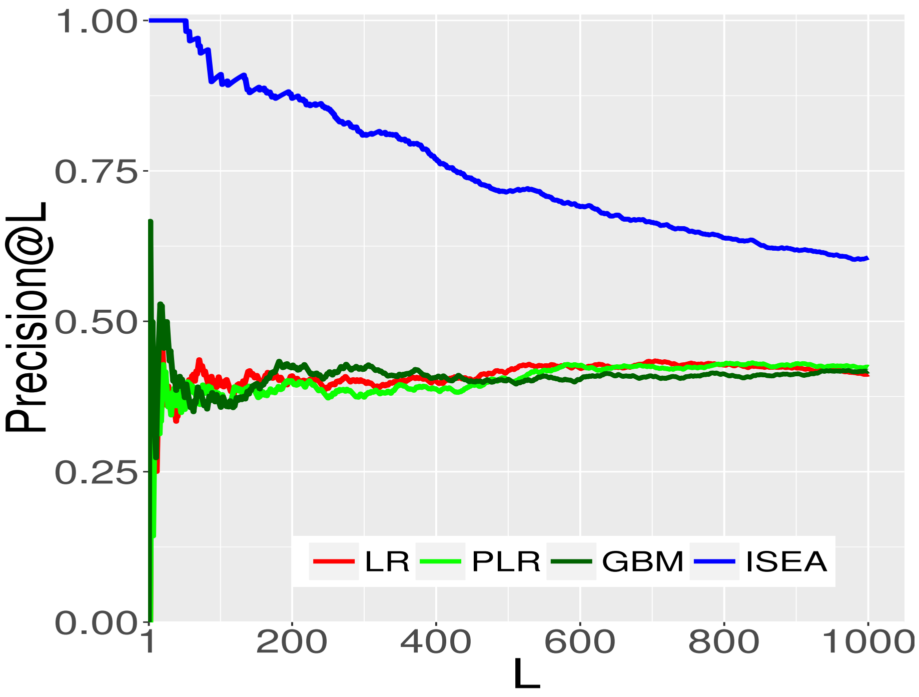


Figure 4.6 Precision over top L predicted records.

As the ultimate goal is to deploy the proposed model in the PPTV video click forecasting platform, our future efforts will be dedicated to two aspects: (1) Admittedly, item-specific effects can change across time. For example, there might be some different viewing patterns between weekdays and weekends. In this case, it is reasonable to construct the time-inhomogeneous item-specific effects. (2) The current evaluation is an offline experiment. The corresponding collection of online click prediction for users will be performed so as to validate and further adjust the model accordingly.

The drawback of ISEA is that the insufficiency of observed click behavior for each video will definitely undermine effects discovery. An extreme example is that it

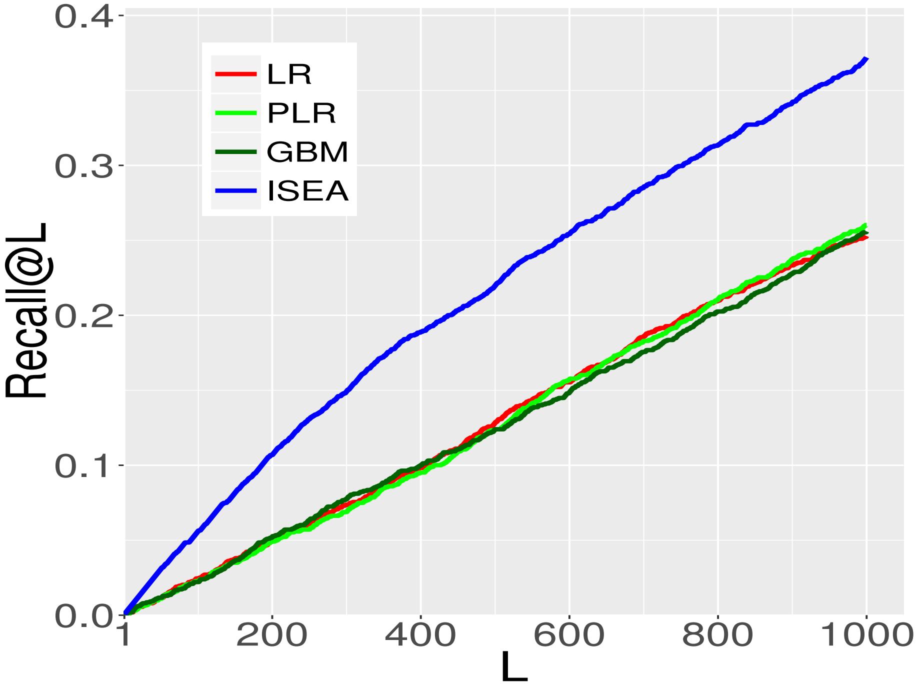


Figure 4.7 Recall over top L predicted records.

fails to estimate item-specific effects of newly added videos. Therefore, one feature work is to further leverage linkage mining to capture the potential interaction among newly added videos and known videos to approximate item-specific effects for new ones [157].

4.6 Conclusion

In order to alleviate the issue caused by the insufficiency of features in video click, we hypothesize the existence of the intrinsic item-specific effects in hidden features. We then present a simple yet powerful ISEA model under the regression framework. The thorough simulation studies are performed to explicate the property and role of the hypothesized effects. Experimental results of PPTV click logs further confirm the existence of item-specific effects in hidden features. They also indicate that the rescued effects can improve the prediction performance against the current algorithms significantly. Our work might probably provide an alternative perspective of rethinking current learning paradigms in terms of feature representation.

CHAPTER 5

BLENDED LEARNING FOR PREDICTING USER INTENDED ACTIONS

5.1 Introduction

User intended actions are widely seen in many areas. Forecasting these actions and taking proactive measures to optimize business outcome is a crucial step towards sustaining the steady business growth. In this chapter, we focus on attrition prediction, which is one of typical user intended actions. Conventional attrition predictive modeling strategies suffer a few inherent drawbacks. To overcome these limitations, we propose a novel end-to-end learning scheme to keep track of the evolution of attrition patterns for the predictive modeling. It integrates user activity logs, dynamic and static user profiles based on multi-path learning. It exploits historical user records by establishing a decaying multi-snapshot technique. And finally it employs the precedent user intentions via guiding them to the subsequent learning procedure. As a result, it addresses all disadvantages of conventional methods. We evaluate our methodology on two public data repositories and one private user usage dataset provided by Adobe Creative Cloud. The extensive experiments demonstrate that it can offer the appealing performance in comparison with several existing approaches as rated by different popular metrics. Furthermore, we introduce an advanced interpretation and visualization strategy to effectively characterize the periodicity of user activity logs. It can help to pinpoint important factors that are critical to user attrition and retention and thus suggests actionable improvement targets for business practice. Our work will provide useful insights into the prediction and elucidation of other user intended actions as well.

5.2 Related Work

In the past decade, the attrition modeling has been widely studied [161]. Numerous works revolved around on binary classification algorithms. The main approach is to build a set of features for users and then train a classifier for the task. Classical data mining algorithms including logistic regression [125], support vector machine (SVM) [36, 49] and random forest [36, 121, 171] are intensively studied for attrition prediction. Among them, random forest is found to be able to achieve the best performance in many fields like the newspaper subscription [36]. Actually, random forest is also the modeling algorithm for customer behavior analysis including attrition or retention behind predictive analytics startup *Framed Data*[1] (acquired by Square) [149]. Besides, some biologically inspired methods like genetic programming [83], evolutionary learning algorithm [7] and vanilla deep neural networks (DNN) [119, 141, 160] were also proposed to search for attrition patterns. Amongst algorithms of this kind, DNN becomes a rapidly growing research direction [141,149,160]. With the growing popularity of deep learning, some advanced methods like convolutional neural networks (CNN) [166] and recurrent neural networks (RNNs) [89] have been utilized recently as well. These works, however, focus on the provided latest attrition status of users. Essentially, they leave out the evolution of historical states inadvertently. The

[1] https://wefunder.com/framed,http://framed.io

precedent statuses would probably be informative in the inference of future statuses by coordinating the feature representation better.

There are sporadic works that have been proposed to exploit historical statuses for attrition prediction [148, 167]. Although these works have tried to incorporate historical statuses of users, they have two issues. First, the whole historical observation periods are divided into multiple sub-periods for model training with handcrafted efforts. In this case, the correlation across different sub-periods cannot be fully explored. Second, the decaying impact of statuses within different sub-periods on attrition prediction within the target time period is out of consideration. The survival analysis framework has been proposed to capture the time-to-event of attrition [103]. It utilizes the initial information at the start of the user enrollment to perform model learning for predicting survival time of subscriptions. Here the inherent problem is that the evolving user activities are not incorporated into the attrition prediction, which are crucial to the attrition modeling according to our experiments. Aside from the above works purely based on attrition, profit-driven discussion and simulation studies were also performed based on a potential intervention assumption (e.g., bonus, discount) [119].

Compared with the intensive research on predictive modeling, little work focuses on the interpretation of attrition prediction results in terms of at both individual and class/group level. This is in part due to inherent challenges faced by non-interpretable classifiers under the framework of traditional interpretation methods [121, 133]. Recently, advanced interpretation methods like saliency maps [145] and its follow-up work Local Interpretable Model-Agnostic Explanations (LIME) [134] have been proposed in this regard. Our technical approach to distilling attrition insights is inspired by saliency maps.

5.3 Modeling

5.3.1 Preliminary

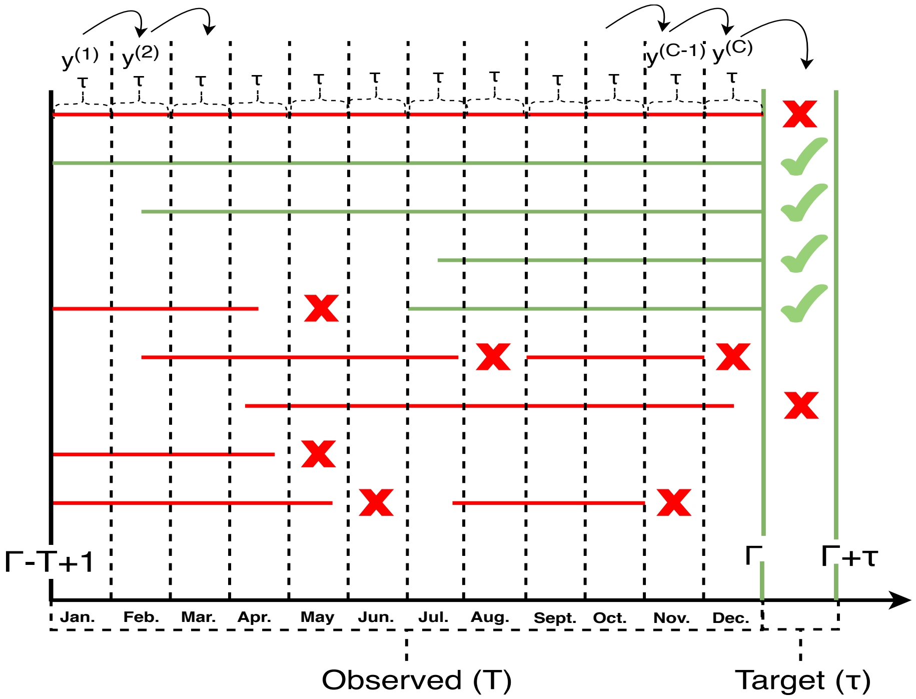


Figure 5.1 Schematic overview of different user statuses with varied types of observed activity logs. There are $C = T/\tau$ snapshots. $\times$ and $\checkmark$ denote attrition and retention, respectively. $\curvearrowright$ indicates that the ground-truth label of user activities during snapshot $t-1$ is the attrition status within future snapshot t, $2 \leq t \leq C$.

In this section, we focus on formulating the attrition prediction problem. To facilitate the problem formulation, we give a schematic illustration of user statuses as shown in Figure 5.1.

Suppose there are a set of N samples or users $\mathbb{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^N$, for which we collect user data $\mathbf{X}_i$ from the historical time period [Γ-T+1, Γ] of length T, and we aim to predict their statuses $y_i \in \mathcal{Y} = \{0, 1\}$ in the future target time window [Γ+1, Γ+τ] of length τ[2].

The user data $\mathbf{X}_i$ are composed of three primitive heterogeneous sub-components: activity logs on a basis of observed time granularity (e.g., day) $\mathbf{X}_{ia}$, dynamic user information $\mathbf{X}_{id}$, and static user profiles $\mathbf{X}_{is}$, namely, $\mathbf{X}_i = (\mathbf{X}_{ia}, \mathbf{X}_{id}, \mathbf{X}_{is}) \in \mathcal{X}$. For the activity logs component, we denote any events happening during the time span T right prior to the target time window as $\mathbf{X}_{ia} = \left(\mathbf{X}_{ia}^{(\Gamma-T+1)}, \mathbf{X}_{ia}^{(\Gamma-T+2)}, \ldots, \mathbf{X}_{ia}^{(\Gamma-1)}, \mathbf{X}_{ia}^{(\Gamma)}\right)$.

The general goal is to search for a reasonable mapping rule from observed feature space to attrition statuses $\mathcal{R}(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$ and subsequently apply $\mathcal{R}(\cdot)$ to estimate the statuses of samples in the future. The probability of sample i in attrition can be denoted as

$$p(y_i = 1|X) \tag{5.1}$$

Practically speaking, the ground truth is relatively subject to the future target time window $[\Gamma + 1, \Gamma + \tau]$. Specifically, if a user drops out of a course or is churned within this window, it is then labeled as 1; if the user remains active, then it is labeled as 0. It is worth noting that attrition labels are generated based on the overall statuses of users during the target time window.

[2] Theoretically speaking, τ is flexible and can be any positive integer. In practice, it depends on business scenarios, which can be weekly, biweekly, monthly or longer periods.

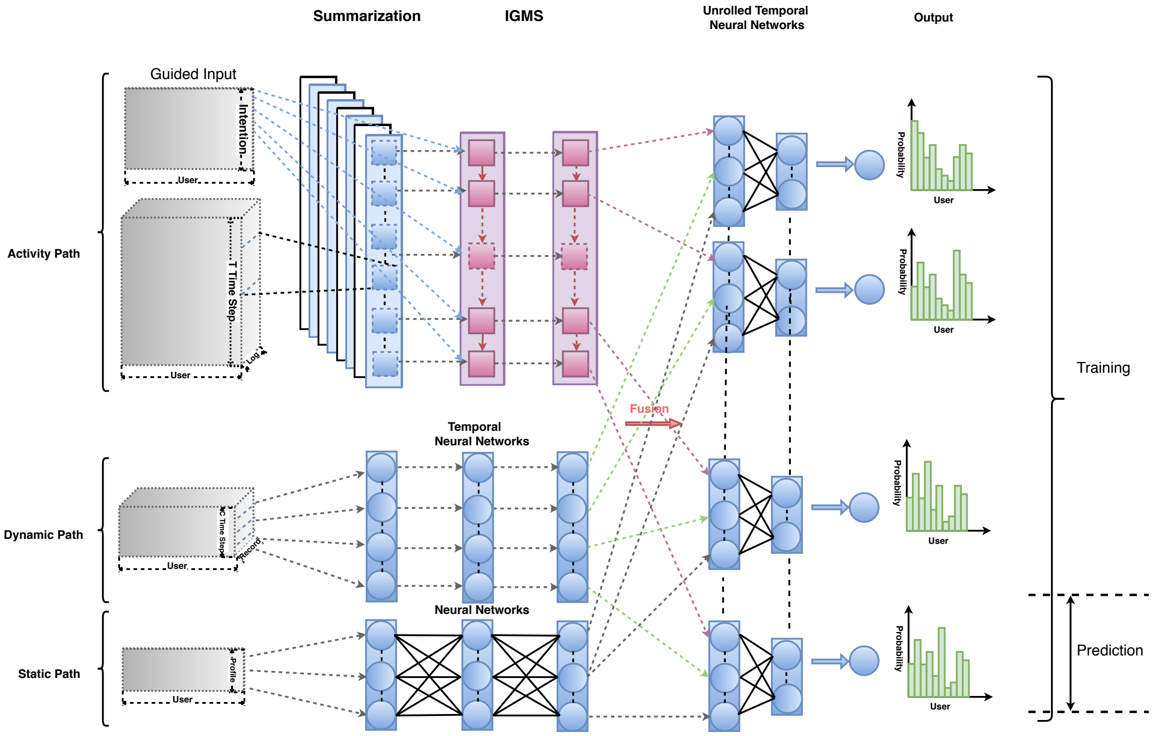


Figure 5.2 Framework overview of Blended Learning Approach.

5.3.2 Methodology

Section 5.3.1 introduces the primitive problem formulation. We propose to extend this formulation to incorporate multi-snapshot statuses according to the snapshot window, which is equal to the pre-designated target time window size τ. Concretely, sequential outputs are generated across sampled observed time period per τ units based on the attrition definition. We then can generate $C = \frac{T}{\tau}$ snapshot outputs. As for users with the observed time span being less than T, we take zero-padding for the computational convenience. The corresponding masking indicators are introduced to disable their contributions to the loss as detailed in Equation 5.8. Accordingly, we obtain the final series of statuses of sample i as $\left(y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(C)}\right)$ where $y_i^{(C)}$ is the status within the target time period. In this case, the conditional probability that sample i is in the state of attrition can be represented as

$$p\left(y_i^{(t)} = 1 | X; y_i^{(t-1)}, \ldots, y_i^{(1)}\right), 2 \leq t \leq C \tag{5.2}$$

Therefore, our learning rule can naturally evolve to be $\mathcal{R}(\cdot) : (\mathcal{X}, \mathcal{Y}^{t-1}) \rightarrow \mathcal{Y}$ for target time step t.

With the reformulation of this problem, we introduce different learning layers/components of BLA and discuss how these components tackle the aforementioned issues faced by the attrition prediction.

Parallel Input Layer In accordance with reformulated mapping rule $\mathcal{R}(\cdot)$, the original feature space includes four different parts: activity logs, dynamic information, static profiles and precedent statuses. In this case, we design multiple parallel input layers for corresponding learning paths to solve the amalgamation problem associated with these heterogeneous multi-view features as diagrammed in Figure 5.2.

Activity input layer– Three-dimensional users activity logs are fed into this layer, along which are user samples, observation time span, and activity metrics. Concretely, the granularity of primitive observation time can be, but not limited to, every minute, hourly, daily, weekly, monthly, or any reasonable time duration. The activities can be, but not limited to, students' engagement for MOOCs, products booting, usage of specific features within the products for software companies.

Dynamic input layer– Three-dimensional dynamic layer is responsible for the derivative of the user profile, products information or their interactive records based on the snapshot window. This includes, but not limited to, subscription age, payment settings (automatic renewal/cancellation), or any reasonable derivatives. ***Static input layer***– This layer takes static profiles of users or products, which cover many details including but not limited to, gender, birthday, geographical location, market segments, registration/enrollment method or any other unchanging information. This layer is of the two-dimensional shape. ***Guided input layer***– The snapshotted statuses as a two-dimensional guided intention is embodied into the attrition prediction through this layer.

Summarization Layer Closely following the activity input layer is the summarization layer, which is developed for summarizing user activities. Due to the homogeneity along the observed time and the heterogeneity across activity logs, we utilize one-dimensional CNN to aggregate low-level activity logs over a fine-grained time span (e.g., day) to generate high-level feature representation over a coarse-grained one (e.g., week). Mathematically speaking, we have

$$f_s^{(t)}(\mathbf{X}_{ib}) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \mathbf{W}_{mn}^{s} \mathbf{X}_{ib}^{(t+m),(n)} \tag{5.3}$$

where $\mathbf{X}$ is the input activity logs, t and s are the indices of output time step and activity summarizer, respectively. Summarizer $\mathbf{W}^s$ is the $M \times N$ weight matrix with M and N being the window size of summarizing time span and sequence channel, respectively. In particular, N of the first summarization layer is equal to the number of activity metrics. Activity logs can be summarized to be with different granularities via setting kernel size M.

The designed summarization layer entails threefold benefits: (1) Learning rich relations and bypassing labor-intensive handcrafted efforts in summarizing primitive activity logs; (2) Upholding the interpretation track of primitive activity metrics compared with hand-operated aggregation; (3) Accelerating the training procedure of model thanks to the noise filtering and feature dimensionality reduction.

Intention Guided LSTM Layer with Multiple Snapshot Outputs In order to capture the long-range interactive dependency of summarized activities and make the most use of generated auxiliary statuses, we propose to introduce a variant of Long Short-Term Memory Networks (LSTM) [79]. To simplify the following notations, we omit sample indices here. The original formulation in the family of Recurrent Neural Networks [136] (RNNs)[3] is usually denoted as

$$h^{(t)} = f\big(h^{(t-1)}, x^{(t)}\big) \tag{5.4}$$

where $x^{(t)}$ and $h^{(t)}$ are the input sequence of interest and the estimated hidden state vector or output at time t, respectively. $h^{(t-1)}$ is the immediate precedent estimated state vector. We here propose to embed the actual immediate precedent status $y^{(t-1)}$ to guide the learning procedure as

$$h^{(t)} = f\big(h^{(t-1)}, x^{(t)}, y^{(t-1)}\big) \tag{5.5}$$

[3] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

As illustrated in Figure 5.3, the core equations are accordingly updated as follows:

$$\begin{aligned}
f_t &= \sigma\Big(W_f\big[h^{(t-1)}, x^{(t)}, y^{(t-1)}\big] + b_f\Big) \\
i_t &= \sigma\Big(W_i\big[h^{(t-1)}, x^{(t)}, y^{(t-1)}\big] + b_i\Big) \\
o_t &= \sigma\Big(W_o\big[h^{(t-1)}, x^{(t)}, y^{(t-1)}\big] + b_o\Big) \\
C_t &= f_t \circ C_{t-1} + i_t \circ \tanh\Big(W_C\big[h^{(t-1)}, x^{(t)}, y^{(t-1)}\big] + b_C\Big) \\
h_t &= o_t \circ \tanh(C_t)
\end{aligned} \tag{5.6}$$

where $\circ$ denotes the element-wise Hadamard product. σ and tanh are sigmoid and hyperbolic tangent activation functions, respectively. f_t, i_t, o_t, C_t are forget, input, output and cell states, which control the update dynamics of the cell and hidden outputs jointly. It is noted that multiple snapshot outputs in the training phase can keep track of the evolution of statuses sequentially and naturally. In the meantime, the introduced auxiliary statuses are complementary to activity, dynamic and static inputs in terms of capturing the intention progression. We call it *IGMS* as annotated in Figure 5.2.

Temporal Neural Network Layer In order to guarantee the temporal order preservation of feature representation, we introduce temporal neural networks:

$$a_l^{(t)} = \sigma(W_l^{(t)} a_{l-1}^{(t)} + b_l^{(t)}) \tag{5.7}$$

where $a_{l-1}^{(t)}$ is a temporal slice of the output of layer $l-1$.

This layer entails twofold roles: (1) Feature learning over different snapshot periods in the dynamic path; (2) Fusion of feature representation in multiple paths.

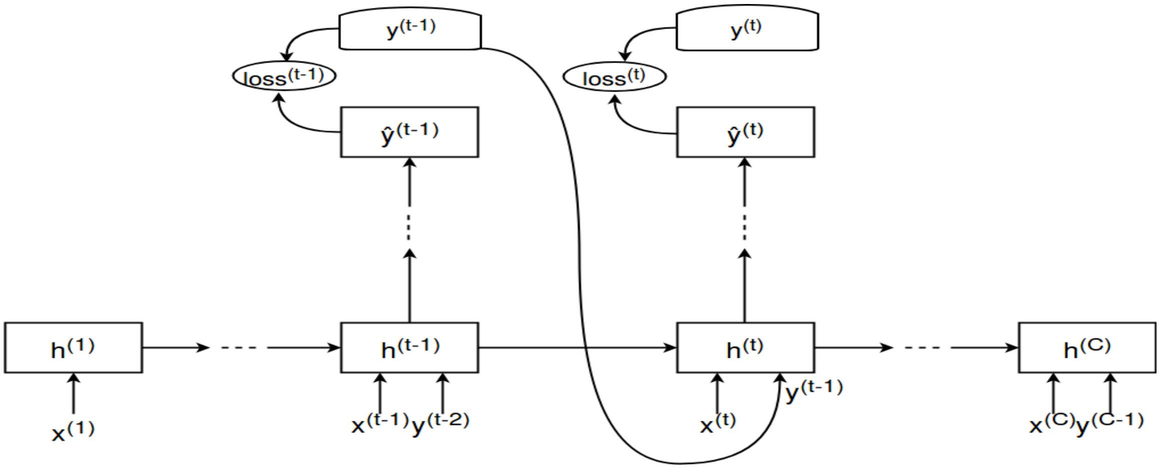


Figure 5.3 Illustration of the guided intention mechanism, where the precedent actual intentions are also used for attrition estimation in next time step.

It is also noted that the activation function of the final temporal neural network layer is *sigmoid*.

Decay Mechanism When multiple snapshot attrition statuses are incorporated into our learning framework, their associated impacts need to be adjusted in the training phase accordingly. This results from the fact that the underlying behavior patterns might change over time in a certain way [152]. To this end, we have a underlying assumption: the bigger the time gap between auxiliary snapshot statuses and attrition status at target time period is, the less similar the underlying intention patterns are. The temporal exponential decay is thus introduced to penalize weights based on this assumption. Concretely, $\zeta^{(C-t)} = k^{(C-t)}$, where $k \leq 1$ depends on the expected speed of decay, as shown in Figure 5.4. Since the decay speed k is a hyper-parameter, it is determined by the validation dataset.

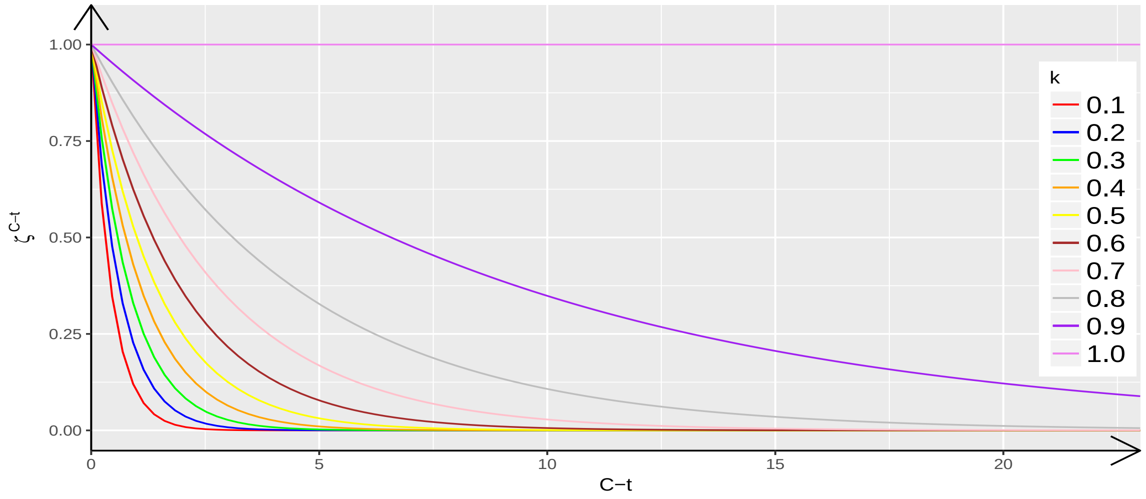


Figure 5.4 Examples of temporal decay of sample weights across different snapshot time steps.

Objective Function With auxiliary snapshot statuses being incorporated into the training phase as shown in Figures 5.2 and 5.3, we have the following loss function to guide the learning procedure:

$$
\begin{aligned}
\mathcal{J} = -\frac{1}{N}\sum_{t=1}^{C}\zeta^{(C-t)}\sum_{i=1}^{N}\eta_i^{(t)}\Big[y_i^{(t)}\log\, p(\hat{y}_i^{(t)} = 1) \\
+ (1 - y_i^{(t)})\log\, p(\hat{y}_i^{(t)} = 0)\Big]
\end{aligned}
\tag{5.8}
$$

where $\zeta^{(t)}$ and $\eta_i^{(t)}$ are temporal decay weight and sample-level binary masking indicator, respectively. In particular, $\eta_i^{(t)}$ can be used to mask invalid attrition statuses of training samples in the snapshot time periods caused by the calendar date alignment. For example, the registration dates of some users are later than the beginning of observed time periods as shown in Figure 5.1.

As shown in Figure 5.2. BLA mainly includes activity path, dynamic path and static path. In the activity path, the time granularity of input is on a basis of primitive observed time granularity (e.g., day), whereas output granularity is the snapshot window (e.g., month). The dynamic path is composed of temporal neural networks with both input and output being at the granularity of the snapshot span. For the static path, the outputs are forked C times for further fusion with outputs of activity and static paths, as shown in unrolled temporal neural networks of Figure 5.2.

Predictive Inference With the learning architecture and estimated parameters, we obtain the learned model ready for predicting user-intended actions. As illustrated in Figure 5.2, we have only one output at the C^{th} time period, which is the attrition probability of the target time period in prediction phase (validation and test).

5.3.3 Feature Interpretation and Visualization

Saliency maps are one powerful technique to interpret and visualize feature representation behind deep neural networks, which have been widely utilized to analyze feature importance [145,155]. In this dissertation, we also construct saliency maps by back-propagating features with the guidance of BLA to highlight how input features impact the user attrition. First of all, a user is supposed to have feature vector x_0 and the associated attrition state, we aim to figure out how elements of x_0 shape output probability of state $\mathcal{R}(x_0)$. Regarding BLA, the score $\mathcal{R}(x)$ is a highly non-linear function of input x. $\mathcal{R}(x)$, however, can be approximated by a linear function in the closeness of x_0 based on the first-order Taylor expansion:

$$\mathcal{R}(x) \approx w^T(x - x_0) + \mathcal{R}(x_0) \tag{5.9}$$

where w is the first-order derivative of $\mathcal{R}(x)$ with respect to the feature vector x at x_0:

$$w = \frac{\partial \mathcal{R}(x)}{x}\Big|_{x_0} \tag{5.10}$$

There are two points about the interpretation of this kind to consider: 1) The magnitude of the derivative indicates the extent to which the change of the most influential elements of feature vector on the probability of the attrition state; 2) The direction of each element of the derivative shows whether such a change boosts or decreases the probability of the attrition state. It is noted that the computation of the user-specific saliency map is very fast due to the requirement of a single back-propagation pass.

For dynamic and static inputs, we take average on saliency maps of all test users and then obtain the overall saliency map. The overall one can help to identify the underlying attrition and retention factors involved in the attrition directly. For activity logs with different metrics, we concentrate on exploring the evolution patterns of activity logs. Thus, we take the absolute value of saliency maps before averaging over all test users. Finally, we take sum of all metrics along observed time periods.

Table 5.1 Basic Statistics of MOOCs and KKBox. Observation Span T and Snapshot Window Size τ are Detailed in Section 5.3.1. τ is Set Based on Business Scenarios (e.g., Subscription Plan) Here.

	# of user	# of attrition	# of persistence	observation span T (day)	snapshot window size τ (day)	target time period
MOOCs	120,542	95,581	24,961	30	10	10 days after the end of observed days
KKBox	11,2118	19,415	92,703	720	30	02/01/2017 ~ 02/28/2017
	156,029	21,752	134,277	720	30	03/01/2017 ~ 03/31/2017

5.4 Experiment

In this section, we first assess the performance of BLA on the customer attrition task comparing with competitive baselines for two public datasets and one private dataset. Then, we perform feature analysis to distill the evolving patterns of user activity logs, attrition and retention factors.

5.4.1 Experimental Setup

We utilize python libraries Keras[4] to build the architecture of our learning algorithm and Tensorflow [1] to perform feature interpretation and visualization. NVIDIA Tesla K80 GPU with memory of 12GB is used for model development. Microsoft Azure with PySpark is adopted as the large-scale data processing platform.

Network Architecture. Along *Activity Path* are 1 one-dimensional CNN (14 kernels) and 2 intention-guided LSTM (30 and 15 kernels). *Dynamic Path* consists of 1 two-layered temporal neural networks with 30 and 15 hidden nodes. *Static Path* involves 1 two-layered neural networks with 30 and 15 hidden nodes. The fusion layer includes 1 two-layered temporal neural networks with 30 and 15 hidden nodes.

[4] https://github.com/keras-team/keras

Training. We initialize parameters in BLA with *Glorot* uniform distribution [64]. The mini-batch size and the maximum number of epochs are set to be 128 and 500, respectively. The parameters are updated based on Adam optimization algorithm [90] with learning rate of 0.001 and decay factor of 1e-3. Early stopping of 20 epochs is set to prevent the overfitting. Trainable parameters and hyper-parameters are tuned based on the loss of attrition records in the validation dataset. As shown in Figure 5.2, all historical records are incorporated into the loss function $\mathcal{J}$ in formula (5.8).

Test. As shown in Figure 5.2, the prediction is conducted on customer attrition records during the target time periods. With both the trained parameters and hyper-parameters, we measure the performance of the model on the specified target periods.

Training and test parts are split based on the temporal logic and will be detailed in the coming subsections accordingly.

5.4.2 Baseline Approaches

In this section, we will introduce alternative algorithms as baseline schemes to demonstrate the effectiveness of the proposed BLA. User activity logs are manually aggregated and then reshaped to be a vector. The one-month dynamic and static information are directly reshaped and then fused with logs vector to generate the learning features. The baselines are tuned based on the validation part and the optimal parameters are reported accordingly.

1. **LR**: The classical Logistic Regression [125] is commonly used with good interpretation capacity [121]. To facilitate the training with the large-scale

dataset, we construct a simple neural network with one input layer and *sigmoid* activation function with GPU acceleration. Adam [90] with learning rate of 0.01 and decay rate of 10^{-3} is adopted as the optimization algorithm for MOOCs and KKBox.

2. **DNN**: Generally speaking, stacking computational units can represent any probability distributions in a certain configured way [15]. Thus, vanilla deep neural networks are widely utilized for attrition prediction in the academic research [141, 160]. The network is with 2 hidden layers of 100, 10 nodes, respectively. Adam [90] with learning rate of 0.01 for MOOCs and 0.001 for KKBox, as well as decay rate of 10^{-3} for both datasets is adopted.

3. **RF**: Random Forest is frequently used in churn or dropout prediction [36, 121, 171] and deployed in the industry (e.g., Framed Data), which usually shows a good performance [149]. The RandomForest of Scikit-Learn with the tree number of 20 and the maximum depth of 30 is employed for MOOCs and KKBox.

4. **NB**: Naive Bayes [81, 122, 180] is also explored in the user attrition prediction. We adopt the classical Gaussian Naive Bayes algorithm for the classification.

5. **SVM**: SVM is explored in this regard as well [36, 49]. To scale better to large numbers of samples (the inherent problem in SVM training), we adopt liblinear (LinearSVC) for *linear* kernel, the bagging classifier (BaggingClassifier + SVC) for non-linear *radial basis function (rbf)* and *polynomial (poly)* kernels in Scikit-Learn library. The settings are linear kernel with $C = 0.001$ for MOOCs and rbf kernel with $C = 0.001$ for KKBox.

6. **CNN**: Convolutional neural networks [166] include two layers of one-dimensional convolutional neural networks with 14 and 7 kernels and the subsequent fully connected neural networks 30 and 15 hidden nodes.
7. **LSTM**: Vanilla recurrent neural networks or long short-term memory networks [50, 89] are also utilized here by aggregating activity logs with handcrafted efforts. One two-layered LSTM with 30-dimensional and 15-dimensional output nodes, followed by subsequent fully connected neural networks with 30 and 15 hidden nodes.

The variants of BLA are listed as follows: *MSMP*: a variant without intention guidance; *IGMP*: a variant without multi-snapshot mechanism; *IGMS-AD*: a variant only using activity path and dynamic path; *IGMS-AS*: a variant only using activity path and static path; *IGMS-DS*: a variant only using dynamic path and static path.

5.4.3 Evaluation Metrics

To measure the prediction performance of the proposed methodology, we adopt F1 Score, Matthews correlation coefficient (MCC) [113], the Area under Curves of Receiver Operating Characteristic (AUC@ROC) [151, 154, 157] and Curves of Precision-Recall (AUC@PR), respectively. As opposed to ROC curves, Precision-Recall curves are more sensitive to capture the subtle and informative evolution of algorithm's performance. A more in-depth discussion is detailed in Ref. [39].

5.4.4 Experimental Results

We perform attrition prediction on two public attrition repositories MOOCs (dropout prediction)[5] and KKBox (churn prediction)[6] based on BLA against baseline approaches. Furthermore, we apply the proposed method to users of Adobe Creative Cloud (CC) and compare it with random forest and the currently deployed model.

MOOCs and KKBox Dropout in MOOCs and churn in subscription-based commercial products or services are two typical scenarios associated with the attrition problem. Dropout prediction focuses on the problem where we prioritize students who are likely to persist or drop out of a course, which is usually characterized by the highly skewed dominance of dropout over persistence. As opposed to dropout in MOOCs, churned users are in tiny proportion compared with persistent ones. The basic statistics of the MOOCs and KKBox datasets are described in Table 5.1 briefly. Here attrition labels indicate the user status within the target time period. As the given spans of the target time period are 10 days for MOOCs and one month for KKBox[7], we set snapshot span as $\tau = 10$ and $\tau = 30$ accordingly. Given observation span $T = 30$ and $T = 720$, a total of $C = 3$ and $C = 24$ outputs are generated simultaneously. The last one is the status to predict, and the precedent outputs are auxiliary statuses for aiding in the model development. User activity logs, dynamic and static features are given in Tables 5.2 and 5.3, respectively. For MOOCs, the stratified data splitting is adopted since there are few overlapping time spans among

[5] https://biendata.com/competition/kddcup2015/

[6] https://www.kaggle.com/c/kkbox-churn-prediction-challenge

[7] τ is pre-specified in datasets here.

different courses. Accordingly, the ratio of training, validation and testing datasets is 6:2:2. For KKBox, user records on Feb 2017 and March 2017 are utilized as model development and assessment, respectively. The development dataset is further split into internally stratified training and validation parts with ratio 8:2.

The comparison results among BLA and baselines are examined in Tables 5.4 and 5.5. Overall, BLA is able to outperform other commonly used methods for attrition prediction in terms of an array of metrics. In MOOCs, we report F1 score and AUC@PR based on minor persistent users, which is sensitive to the improvement of algorithms. It is noted that, as compared to baselines, the performance gain of BLA is more obvious in KKBox than that in MOOCs. There are two underlying causes: (1) Few dynamic and static user features are available for MOOCs, which degrades the power of the multi-path learning as shown in Table 5.3; (2) The span of historical records is limited for MOOCs, which will suppress the multi-snapshot mechanism inevitably as shown in Table 5.1.

Ablation Analysis To explore the potential explanation for BLA's performance, a series of ablation experiments are conducted to study the role of key components of BLA. We focus on KKBox here due to the limited observation time span of MOOCs.

First of all, we empirically study the decay mechanism. As shown in Figure 5.4, $k = 1$ indicates all auxiliary statuses share equivalent weights in loss function, which also means that no decay mechanism is considered here. Meanwhile, when $k \to 0$, it implies that auxiliary snapshot statuses are ignored and only the status of target time period is considered. The quasi U-shaped curve of loss on validation dataset demonstrates the existence of decay in attrition patterns over observed time steps

Table 5.2 Student Engagement Logs of MOOCs and Customer Music Listening Logs of KKBox.

Dataset	Activity	Remarks
MOOCs	problem	working on course assignments
	video	watching course videos
	access	accessing other course objects except videos and assignments
	wiki	accessing the course wiki
	discussion	accessing the course forum
	navigate	navigating to another part of the course
	page_close	closing the web page
	source	Event source (server or browser)
	category	the category of the course module
KKBox	num_25	# of songs played less than 25% of the song length
	num_50	# of songs played between 25% to 50% of the song length
	num_75	# of songs played between 50% to 75% of the song length
	num_985	# of songs played between 75% to 98.5% of the song length
	num_100	# of songs played over 98.5% of the song length
	num_unq	# of unique songs played
	total_secs	total seconds played

as presented in Figure 5.5. Furthermore, the evidence that the value of right side is less than that of left side suggests the necessity of the proposed multi-snapshot strategy. The performance of different variants of BLA is also reported in Figure 5.6. Their performance disparity delivers useful points. To be specific, it is activity path

Table 5.3 User Dynamic and Static Profile of MOOCs and KKBox.

Dataset	Style	Profile	Remarks
MOOCs	dynamic	—	—
	static	course_id	course ID
KKBox	dynamic	membership	the time to the initial registration
		is_auto_renew	whether subscription plan is renewed automatically
		is_cancel	whether subscription plan is canceled
	static	bd	age when registered
		city	city when registration (21 anonymous categories)
		gender	gender (male and female)
		registered_via	registration method (5 anonymous categories)

Table 5.4 Performance Comparison on MOOCs for Attrition Prediction

Method	**AUC@ROC**	**MCC**	**F1 Score**	**AUC@PR**
LR	0.8595	0.5477	0.6017	0.7003
RF	0.8693	0.5753	0.6430	0.7077
DNN	0.8718	0.5786	0.6509	0.7157
NB	0.8354	0.4976	0.5925	0.6562
SVM	0.8656	0.5440	0.5924	0.7049
CNN	0.8778	0.5851	0.6480	0.7324
LSTM	0.8746	0.5863	0.6528	0.7246
BLA	**0.8842**	**0.5973**	**0.6569**	**0.7464**

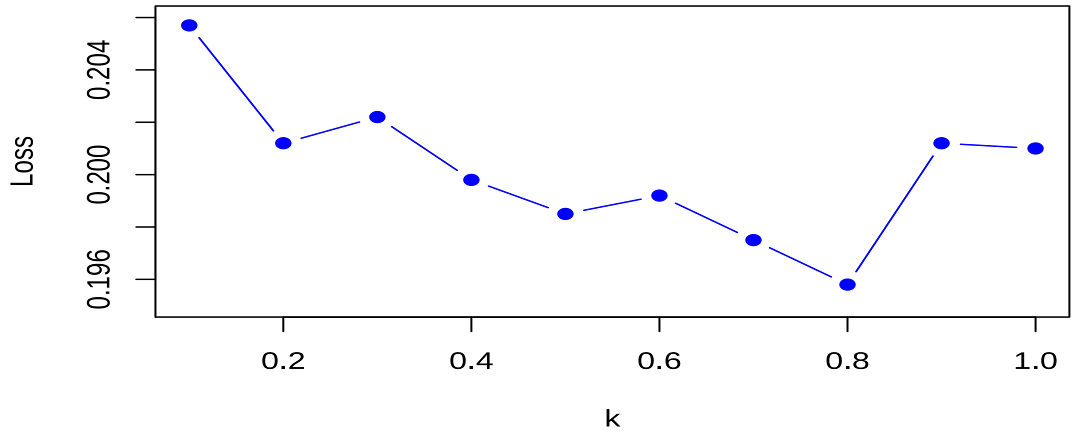


Figure 5.5 The evolution of validation loss over decay speed k for the exponential decay.

Table 5.5 Performance Comparison on KKBox for Attrition Prediction

Method	**AUC@ROC**	**MCC**	**F1 Score**	**AUC@PR**
LR	0.8292	0.6560	0.6986	0.7360
DNN	0.9016	0.6005	0.6552	0.5937
RF	0.9394	0.6961	0.7314	0.8085
NB	0.5061	0.0314	0.2467	0.5671
SVM	0.5960	0.2111	0.1091	0.4972
CNN	0.8963	0.5409	0.5974	0.5263
LSTM	0.9293	0.6786	0.7171	0.7779
BLA	**0.9600**	**0.7280**	**0.7621**	**0.8436**

that has the highest impact, followed by dynamic path and finally static path. Both intention guidance and multi-snapshot mechanisms shape BLA in different manners as well.

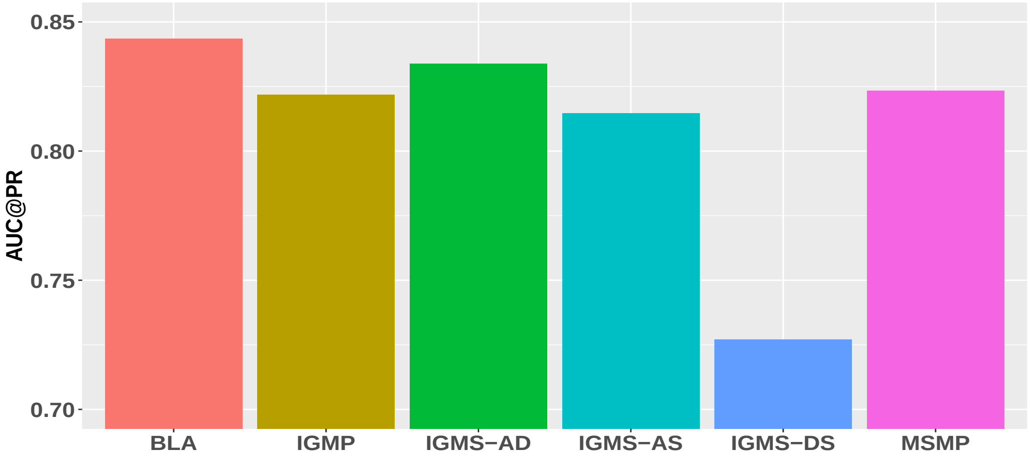


Figure 5.6 Performance comparison for different variants of BLA. Variants are described in Section 5.4.2.

Attrition and Retention Factors After attrition prediction, the next step typically is to identify underlying patterns/indicators or to explore feature importance.

Regarding user activity logs, the feature importance across different observed time steps are visualized in Figure 5.7. Overall, the feature importance changes periodically with the peak value in the vicinity of the intersection of two successive snapshots. Furthermore, the peak increases roughly as observation moves onwards to the target time period. The locality of peak values indicates user activities around payments are informative and important compared with other time steps. The evolution of peak values across different time periods shows that attrition within

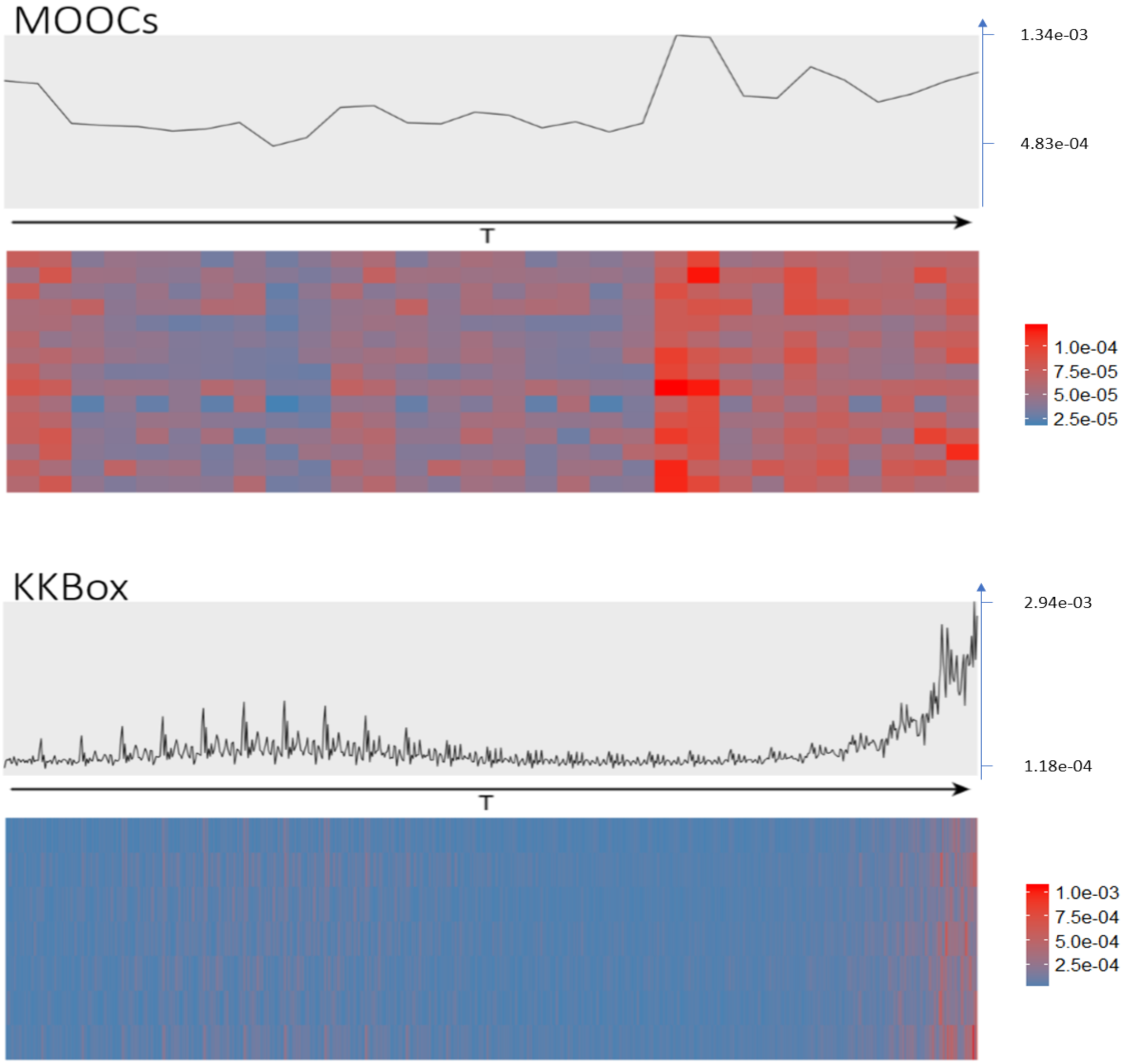


Figure 5.7 Heat map of feature importance of the MOOCs and KKBox datasets.

target time period is highly related to the proximate user activities, which is also intuitively reasonable.

When it comes to dynamic and static user information, we also do in-depth analysis on KKBox. Among them, the most important features are *is_cancel*, and *is_auto_renew* from the dynamic side, and *registered_via* from the static side. As the registered method is provided anonymously, we cannot do any explanation. As shown in the top of Figure 5.8, field *is_cancel* indicates whether a user actively cancels a subscription or not, which is proven to be positively correlated with attrition. It might be due to the change of service plans or other reasons, though. Naturally, feature *is_auto_renew* shows the intention of users to persist, which is also confirmed by the negative saliency value.

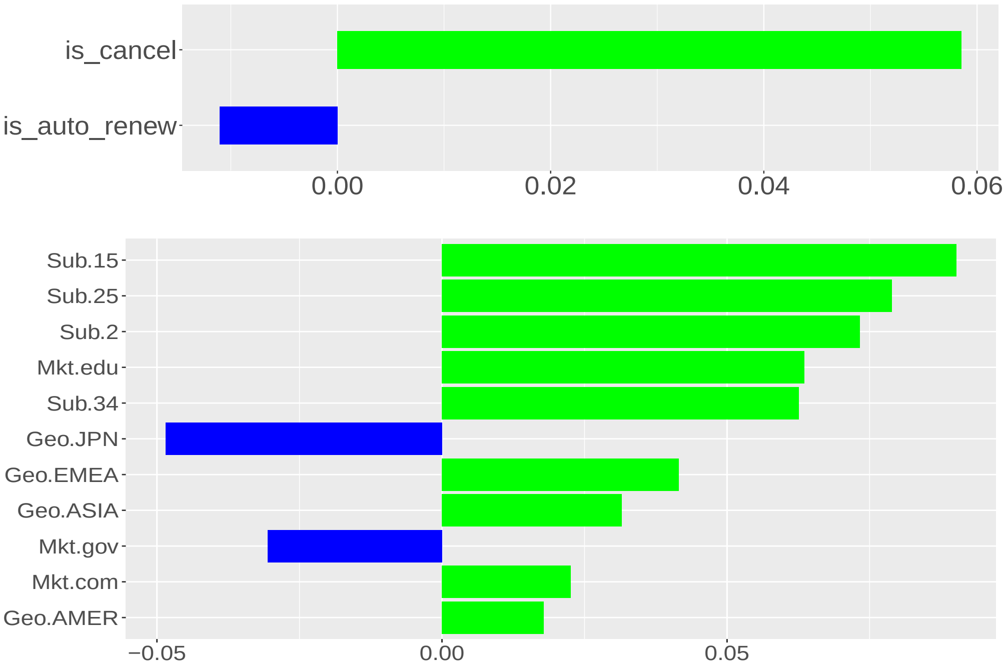


Figure 5.8 Attrition and retention factors for KKBox (top) and Adobe CC (bottom). Features are detailed in Table 5.7.

Adobe Creative Cloud Adobe CC provides entire collection of desktop and mobile applications for the brilliant design, which is characterized by low user attrition rate. We apply the preliminary version of BLA (without decay mechanism or guided intention) called *pBLA*[8] to perform churn prediction and analysis on sampled users, which are briefed in Table 5.6. Concretely, user activity, dynamic and static information used in our model are described in Table 5.7. Regarding activity logs, two daily metrics booting times and total session time for each application (e.g., Photoshop) are recorded. Besides, we conduct both monthly and annual discretization of subscription age to capture two representative subscription types adopted by Adobe CC.

In our experiments, we consider users with subscription age of within 3 years. Due to confidentiality restrictions, we cannot disclose the volume of attrition and retention users. The dataset with the target time period of May 2017 is used for model development in which churned, and persistent users are sampled equivalently. We then evaluate the predictive capacity of our algorithm in two scenarios. In the first scenario, the test dataset includes sampled users with a ratio of 1:1 during target time period of June 1 to June 30, 2017. We then compare pBLA with widely used random

[8] Decay mechanism was not considered into our model during the internship period yet.

Table 5.6 Basic Statistics of Adobe CC Users As of the Beginning of the Target Time Period. Observation Span T (Day), Snapshot Window Size τ (Day)

sampling	$\frac{\text{\# of attrition}}{\text{\# of persistence}}$	T	τ	target time period
yes	1	360	30	05/01/2017 ∼ 05/31/2017
yes	1	360	30	06/01/2017 ∼ 06/30/2017
no	—	360	30	07/01/2017 ∼ 07/31/2017

Table 5.7 Activity Logs of Applications, Dynamic and Static Information for Adobe CC Users

	Feature	Remarks
Activity	Ps	booting times and total session time of Photoshop
	Ai	booting times and total session time of Illustrator
	Id	booting times and total session time of InDesign
	Pr	booting times and total session time of PremierePro
	Lr	booting times and total session time of Lightroom
	Ae	booting times and total session time of AfterEffects
	En	booting times and total session time of MediaEncoder
Dynamic	Sub	the subscription age of Adobe CC
Static	Mkt	market segment (education, government, and commercial)
	Geo	general geographical code (JPN, EMEA, ASIA, AMER)

forest in the industry (e.g., Framed Data) for attrition prediction [36, 149, 171]. The results are reported in Figure 5.9. The significant performance gain can be gained here. In the other scenario, we compare our model with currently deployed model[9] on users who were still active at the end of June 2017 (without sampling). Our proposed model beats the currently deployed model greatly as reported in Figure 5.10[10]. The superiority of our algorithm over other approaches is more evident in

[9] Features are created based on user profile and products usage logs. For the product usage feature, we mainly utilized user usage records of 7 top Adobe CC products to generate counts, rates, recency over different time windows for different types of events. We also performed extensive feature engineering, such as imputation, capping, logarithm, binning, interactions of two variables like ratios and products. Logistic regression based on multi-snapshot data was trained with elastic net regularization. Model hyper-parameters are tuned based on 5-fold cross-validation with the best of efforts.

[10] The attrition probability adjustment of the currently deployed model is based on all users beyond the subscribed age of 3 years. We thus omit threshold based evaluation metrics.

Adobe CC than that in other datasets. This mainly results from the difference of the subscription plans. Most subscriptions of Adobe CC are the type of annual plan while other datasets experience a couple of months (e.g., 30 to 90 days for most KKBox subscriptions). The evolution of intended actions across long subscription plan period is amenable to our algorithm.

Likewise, the feature analysis implies that activity logs of users on applications of Adobe CC are characterized by the explicit periodicity in terms of impacts on attrition as shown in Figure 5.11. Due to the long subscription plan for Adobe CC as mentioned before, the maximum of periodical peak values might be earlier than within the last month. Additionally, as shown in the bottom of Figure 5.8, subscription age plays a very import role, for example, 15^{th}, 25^{th}, 2^{nd}, 34^{th} month are the most risk months since the beginning of subscription, which are all around the renewal dates of annual subscription plan[11]. Regarding static information, Japan (JPN) is found to be the most persistent area compared with other geographical areas. Also, it is easy to expect churn in subscribed users for the educational purpose, followed by the commercial and finally governmental purposes.

5.5 Discussion and Future Work

The introduced user alignment based on the calendar timeline enables an unbiased modeling. The multi-path learning helps to fuse multi-view heterogeneous features, and the summarization layer is introduced to aggregate and integrate primitive user activity logs. In addition, we leverage IGMS with decay mechanism to track evolving

[11] Monthly installment payment is available for the annual membership of Adobe CC.

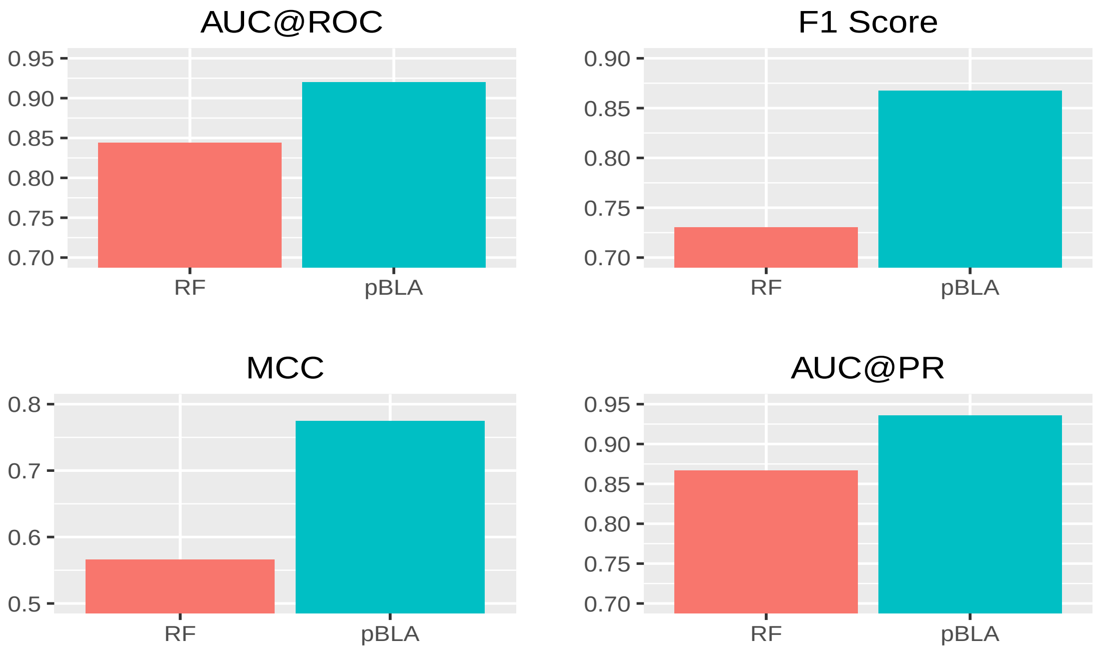


Figure 5.9 Performance comparisons between pBLA and random forest.

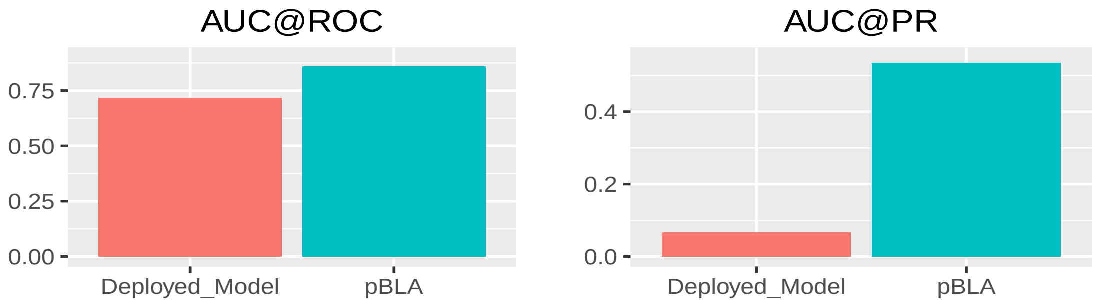


Figure 5.10 Performance comparisons of pBLA against the currently deployed model.

intentions. Finally, saliency maps are introduced to elucidate the activity patterns, attrition and retention factors. There are some interesting aspects to explore in

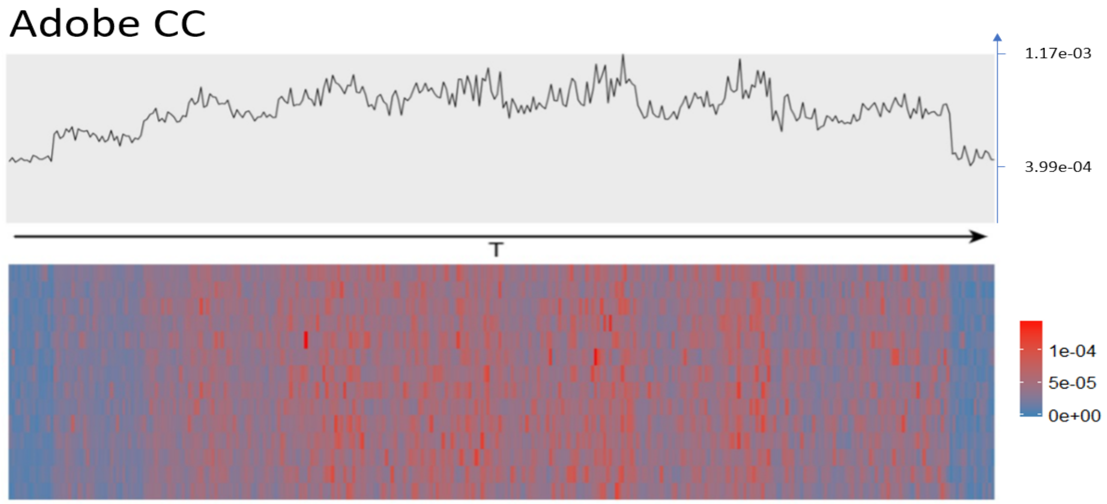


Figure 5.11 Heat map of feature importance for Adobe CC user churn.

the future. First of all, from the perspective of the marketing campaign in the industry, the cost of attrition and retention may not be equivalent under some commercial circumstances. Thus, the probability threshold and corresponding loss function can be adaptively adjusted to account for their business profitability. In this case, some profit-driven strategies can be designed accordingly. Second, we consider the commonly used exponential decay in a trial-and-see way to explore the impacts of different time periods on the status of current time steps of interest. The hyper-parameter k is determined by the validation dataset. It is desirable to develop a principled and feasible way to tune k automatically and even discover the underlying decay evolution involved in the attrition prediction without the distribution assumption. This remains the topic of our future research.

5.6 Conclusion

In this chapter, we explore the classical attrition prediction (dropout and churn) problem and elucidate the underlying patterns. The proposed BLA is able to address an array of inherent difficulties involved in traditional attrition prediction algorithms. Particularly, the exploration of the decay mechanism further demonstrates the power and flexibility of our BLA in terms of capturing the evolving intended actions of users. The extensive experiments are conducted on two public real-world attrition datasets and Adobe Creative cloud user dataset. The corresponding results show that our model can deliver the best performance over alternative methods with high feasibility. The feature analysis pipeline also provides useful insights into attrition. Our work can also be applied to the attrition problem in related areas and other user intended actions.

CHAPTER 6

ELUCIDATION OF DNA METHYLATION ON N^6-ADENINE

6.1 Introduction

We developed a deep learning-based algorithm to predict DNA N6-methyladenosine (6mA) sites de novo from sequence at single-nucleotide resolution, with application to three representative model organisms, A. thaliana, D. melanogaster and E. coli. Extensive experiments demonstrate the accuracy of our algorithm and its superior performance compared with conventional k-mer based approaches. Furthermore, our saliency maps-based context analysis protocol reveals interesting cis-regulatory patterns around the 6mA sites that are missed by conventional motif analysis. Our findings will help to elucidate the regulatory mechanisms of 6mA and benefit to the in-depth exploration of their functional effects. Lastly, we offer a complete catalog of 6mA sites based on in silico whole-genome prediction. Enrichment analysis of this complete catalog shows that 6mA is enriched in tRNA gene regions across different organisms.

6.2 Results

It is worth noting that our proposed prediction is purely based on sequence information where only cis-effect will be captured. Whether a candidate is 6mA site or not will also depend on many other exogenous trans-effects. Therefore, what our method predicts is the candidacy or potential for being a 6mA site. Since we dont aim to make any developmental or tissue specific 6mA prediction, the methylation data

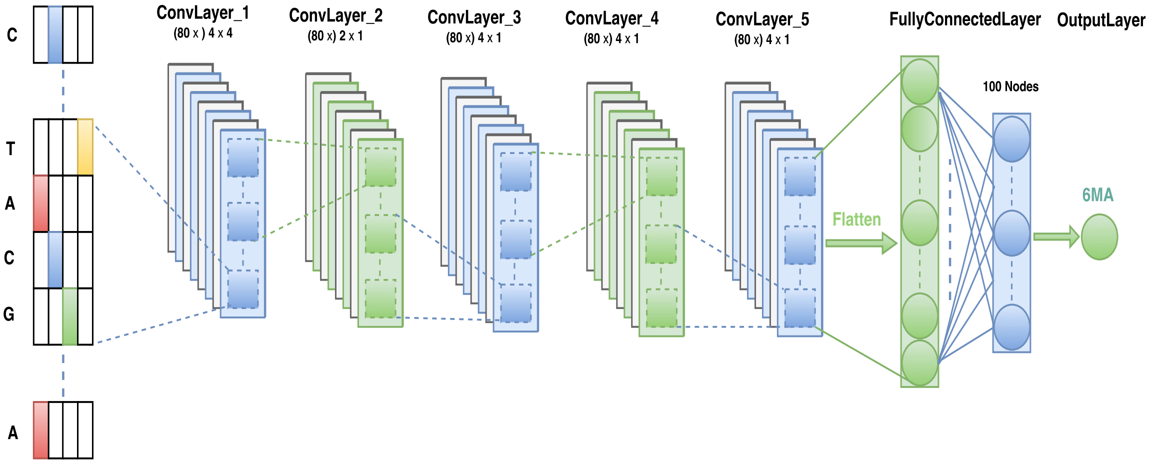


Figure 6.1 The proposed network architecture of DeepM6A, a method for predicting DNA modification on N6-Adenine.

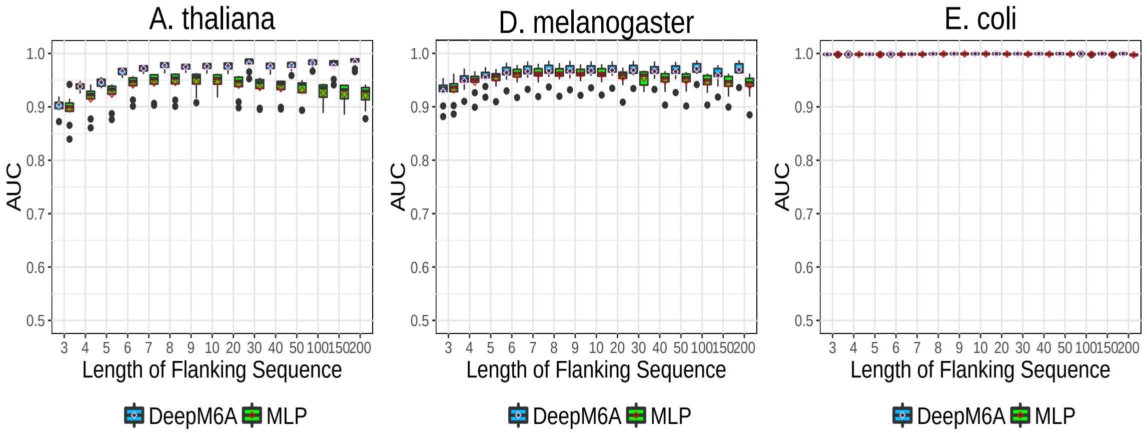


Figure 6.2 Performance comparison of DeepM6A against MLP (standard multi-layer perceptron networks) for different lengths of flanking sequence 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 100, 150, 200 on A. thaliana, D. melanogaster and E. coli in terms of AUC.

we used in our model were collected from a mixture of cells at different development stages. Of our interest is to predict the candidacy or potential for being a 6mA

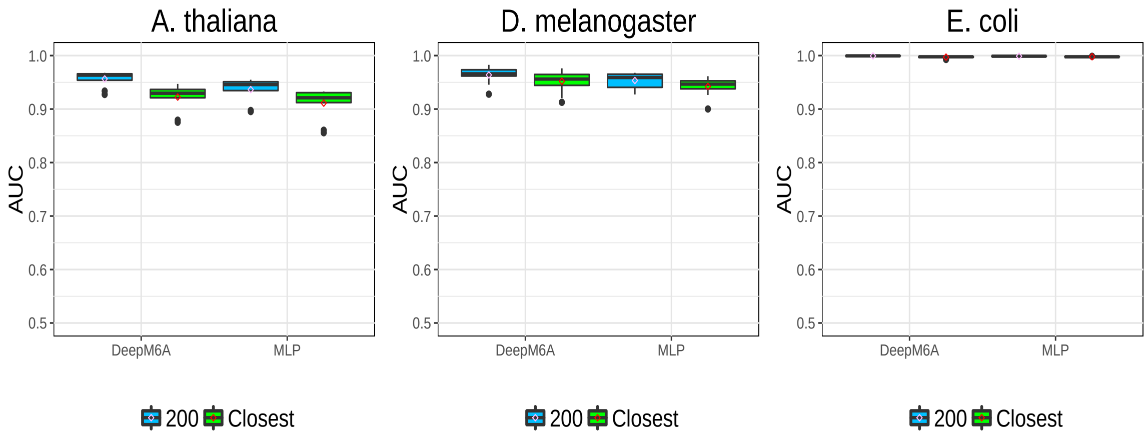


Figure 6.3 Performance comparison of DeepM6A against MLP (standard multi-layer perceptron networks) for different distance scales of methylated and non-methylated cohorts in terms of AUC. The distance scales are set to be 200 and the minimum (closest sites), respectively.

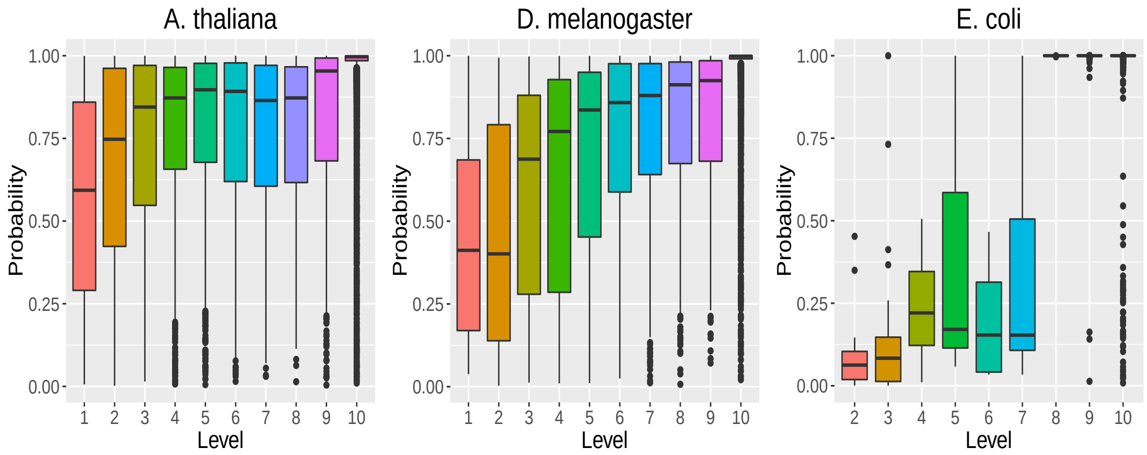


Figure 6.4 Predicted probability versus the methylation fraction/level for A. thaliana, D. melanogaster and E. coli, respectively. The whole 6mAs are separated into 10 levels based on methylation fractions (1: 0 - 0.1; 2: 0.1 0.2; ... ; 9: 0.8 0.9; 10: 0.9 - 1). More than 99% methylated adenine sites locate in methylation level of 9 and 10 for E. coli.

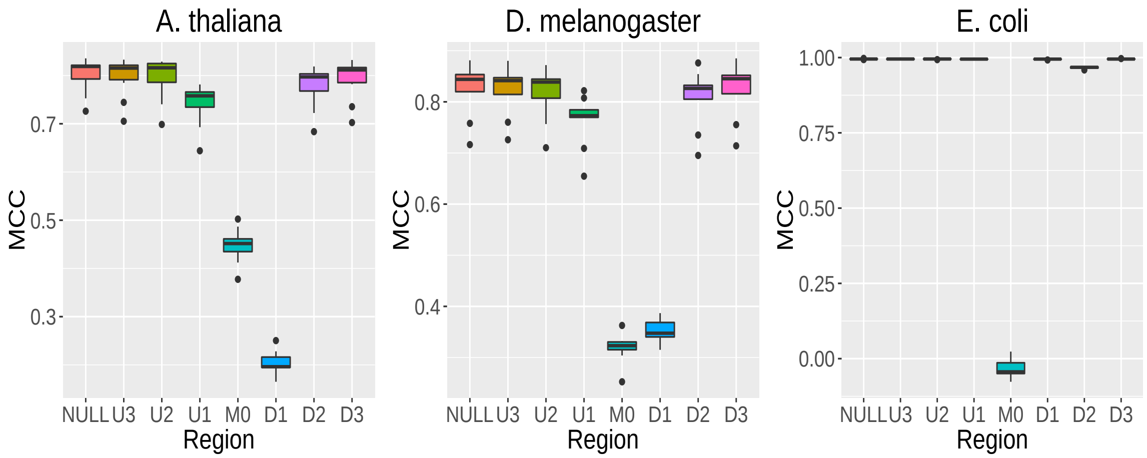


Figure 6.5 Performance comparison of DeepM6A across different perturbed regions [-30, -13] (U3), [-12, -8] (U2), [-7, -3] (U1), [-2, +2] (M0), [+3, +7] (D1), [+8, +12] (D2), [+13, +30](D3) in terms of MCC for A. thaliana, D. melanogaster and E. coli, respectively. Region NULL represents that none of flanking sequences are perturbed.

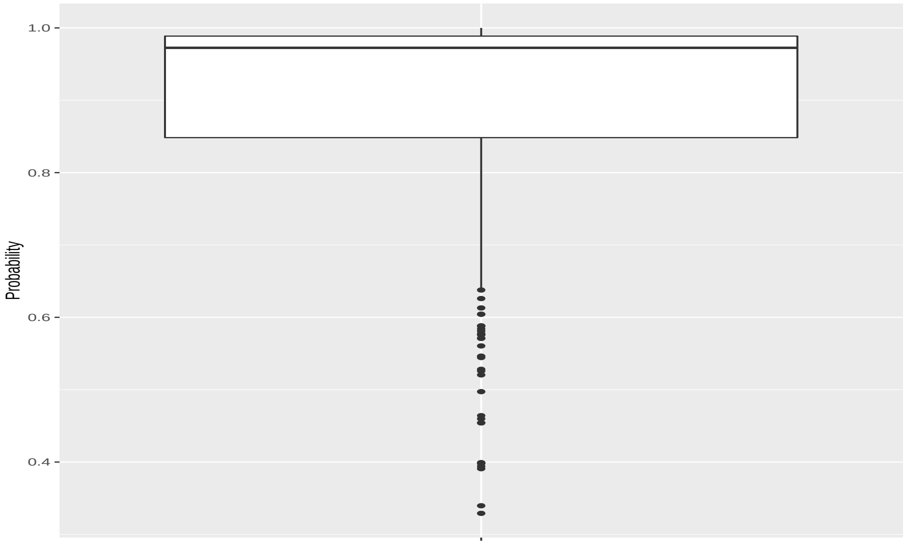


Figure 6.6 DeepM6A-based predictive probability of peak regions called by 6mA-DNA-IP-SEquation.

site, which implies a necessary condition but not a sufficient condition. Such 6mA prediction is quite similar as gene prediction or gene finding in the early Bioinformatics era, which refers to the process of identifying the regions of genomic DNA that encode genes. Most gene prediction methods utilize DNA sequence information only, as our method does.

To predict 6mA candidate sites, we first developed a deep convolutional neural networks-based [5, 92] end-to-end algorithmic framework (Figure 6.1) to capture sophisticated regulatory patterns for predicting 6mA sites de novo from genomic sequences (DeepM6A). Machine learning methods have been employed for genomic sequence-based prediction. Most of them reply on human handcrafted features, e.g., k-mer for predicting mutation effect [93] and polyadenylation sites [84], among many others. Compared with k-mer based methods, our proposed DeepM6A have four major advantages: automating the sequence feature representation of different granularities, hierarchically; integrating a broad spectrum of flanking context sequences, effectively; enabling the potential visualization of inherent sequence motifs for interpretation, naturally; and facilitating model development and prediction in large-scale genomic data, seamlessly. The first two desirable properties jointly contribute to the appealing predictive capacity of DeepM6A. Based on the third property, we then introduced a novel learning protocol to decode the underlying methylation patterns. Both cis-regulatory elements and regions are identified, which will offer useful insights to the in-depth exploration of underlying formulating and regulatory mechanisms of 6mA. Exploiting its accurate prediction, we performed whole-genome scan using DeepM6A for cataloging all potential 6mA sites. Further enrichment analysis of the complete *in silico* 6mA catalogs revealed the enriched distribution of 6mA in tRNA genomic regions. This 6mA-tRNA association lends support to our recent hypothesis that the genomic region of tRNA may form a similar secondary structure as tRNA, which then can be recognized and methylated by the tRNA methyltransferase.

6.2.1 DeepM6A Accurately Predicts 6mA Candidate Sites

We tested DeepM6A in three representative model organisms, namely, A. thaliana (eukaryote, plant), D. melanogaster (eukaryote) and E. coli (prokaryote). As a benchmark, classical k-mer based logistic regression (LR) was also evaluated. The raw SMRT-seq data of A. thaliana, D. melanogaster and E. coli were collected from the PacBio public database. Base modification detection was done to generate an initial set of 6mA sites following the automated data analysis workflows recommended by PacBio. To reduce false positives, we further filtered out the following candidates: (1) any sequence variance located between 10bp upstream and 5bp downstream of the identified modification site; (2) the variation of estimated methylation level is greater than 30%. As a result, we ended up with 19,632, 10,653 and 33,700 6mA sites for A. thaliana, D. melanogaster and E. coli, respectively. These sites account for 0.025696% (76,401,454), 0.013418% (79,393,495) and 1.475402% (2,284,124) proportion of whole-genome adenine sites. The above 6mA sites were used as positive samples in prediction models. To generate negative samples, we sampled the same numbers of non-methylated adenine sites randomly from the whole-genome sequences. At the same time, for the sampled negative non-methylated sites, we required that their distance to any positive methylated site be at least 200 bp away. Then, for both positive and negative samples, contextual sequences around the adenine site at each side were extracted as input for predictive models. We considered the lengths of flanking sequences from 3 bp to 200 bp. We divided all positive and negative samples into three sets for training, validation and testing, respectively, based on their genomic locations. Specifically, for each chromosome of species, we split it into 10 equal segments. We then randomly picked one segment and used the samples within that segment for testing. The samples on the nearest half upstream and half downstream segments were used for validation. The rest of all sites were used for training purpose (Green). In this manner, we had a ratio of 8:1:1 among training, validation and testing datasets, where training and testing parts are strictly non-overlapped. Taking the +/- 30bp flanking sequences as input, DeepM6A is capable of accurately predicting 6mA sites with average area under the receiver operating characteristic curves (AUC) of 0.9564, 0.9637 and 0.9994 for A. thaliana, D. melanogaster and E. coli, respectively (Figure 6.7a), as evaluated by the holdout testing genomic sequences (Online Methods). The salient performance difference between three model organisms implies the more challenging task of identifying 6mA sites and associated sophisticated patterns in advanced eukaryotes than in primitive prokaryotes.

6.2.2 DeepM6A Effectively Exploits Signals from Contextual Sequences

We varied the contextual length of input sequence from 3bp to 200bp, to demonstrate the capability of DeepM6A in exploiting signal information from contextual sequences. With varied input sequence lengths, DeepM6A outperformed LR consistently (Figure

6.7b). Specifically, the performance of DeepM6A keeps improving with the length and reaches a plateau after 10bp for both A. thaliana and D. melanogaster. In contrast, for LR, while increasing the length is beneficial initially, it has opposite effects after 7bp. This finding confirms that the immediate up/down-stream 7-10bp region of 6mA site is critical [185]. However, there may be additional subtle and/or sophisticated signals beyond the 10bp position. This distant signal can be captured by DeepM6A as indicated by its increased performance, whereas the extended region proves deleterious for the k-mer based approach. We attribute the DeepM6A's superiority and robustness to its hierarchical representation of regulatory patterns and accuse the k-mer based methods of their inherent drawbacks of handcrafted feature extraction.

6.2.3 DeepM6A Maintains Good Sensitivity at Single Nucleotide Resolution

The non-N6-methylated adenines in the control cohort are at least 200bp away from any 6mA. To show the robustness of DeepM6A with single-nucleotide sensitivity, for each 6mA site, we selected its closest non-N6-methylated adenine and built a new control cohort ($>$ 75% fall within 5bp of 6mA sites). At the contextual length of 30bp, there is a substantial overlap between cases and the new controls, making separation of 6mA from control more challenging. We also re-evaluated previously trained models. The performance of DeepM6A drops a little, but remains consistently high, while the performance of LR deteriorates substantially (Figure 6.7c). This robustness of DeepM6A advocates its application to 6mA prediction at single-nucleotide resolution.

6.2.4 DeepM6A Outperforms Standard Deep Learning Approaches

In addition to the classical k-mer based LR, we also compared DeepM6A with a standard multi-layer perceptron network [135] (MLP) that uses the same input as our method for predicting N6-methyladenine sites. The input of MLP is also the one-hot encoding of 61 nucleotides centered on the target adenine. The training, validation and testing procedures exactly followed the way DeepM6A was optimized. The predictive capacity under different contextual sequences and the robustness of single nucleotide sensitivity are reported (Figures 6.2 and 6.3). DeepM6A outperforms MLP particularly for longer flanking sequences. As with LR, the performance of MLP keeps improving with the length initially but degrades slightly after 10bp (Figure 6.2). Regarding single nucleotide sensitivity, both DeepM6A and MLP share the similar robustness. Overall, the one-hot encoding is a better feature representation by preserving the primitive sequences in comparison with k-mer format. The hierarchical feature extraction of DeepM6A is more powerful than that of MLP. We, therefore, conclude that DeepM6A is both precise and robust in predicting 6mA. Its superiority is at least in part attributed to its deep network structure, which uses several hidden layers to learn a high-level representation of the DNA sequence hierarchically. To elucidate the power of this hierarchical representation and learning, we visualize the positive and negative samples using t-SNE [109] based on the features learned at different network layers. The features become more and more discriminative along the layer hierarchy, with methylated and non-methylated sites mixed at the input layer, whereas a clear separation culminating in the output layer. Interestingly, with higher methylation level, the better the separation, as observed in the last layer

(Figure 6.8a-c). This is also consistent with the observed high correlation between predicted probability and methylation level (Figure 6.4).

6.2.5 SM-CAP Can Reveal Advanced Cis-regulatory Patterns

After identifying 6mA, the next step typically is to search for regulatory sequences in surrounding regions. Unlike conventional motif analysis [11, 45, 75], we developed a saliency maps-based context analysis protocol (SM-CAP). SM-CAP works by quantifying the contribution of a single base in the modeling context of all other participating bases, such as the non-linear models DeepM6A employs. In contrast, conventional motif analysis assumes simple independence and additive effects among regulatory bases [75]. Our current knowledge suggested that genomic 6mA may not be conserved. Such conservancy status is defined conventionally as traditional motif dentition. Advanced nonlinear motif patterns can exist and exhibit unconventional conservancy status. Take the following simulated data for example.

Suppose there is a 61-bp DNA segment with Adenine (A) in the center, and A, C, G, and T distributed evenly at the rest loci. The central A will get methylated if and only if a combination of A, C, G, T shows at four specific loci X1, X2, X3 and X4. The order A, C, G, and T does not matter, for example, ACGT, TCGA or GATC, can all lead to the methylation of the central A. We can see that the motif pattern at the four loci X1, X2, X3, X4 is not conserved, according to conventional conservancy status definition. Conventional motif search algorithms would fail to recognize this motif. As shown in the following figure, our proposed SM-CAP can successfully identify the four loci and assign appropriate importance scores for the four bases, in comparison with other irrelevant loci.

Our SM-CAP can thus capture advanced patterns that are missed by traditional analysis, as verified by the simulation studies (Online Methods).

6.2.6 Cis-regulatory Patterns of 6mA Revealed by SM-CAP

We used SM-CAP to analyze and visualize the contextual region of 6mA for the three species (Figure 6.9a-c). We can see that the central region is the most critical and that eukaryotes exhibit more sophisticated patterns than the prokaryote. Interestingly, we observe asymmetrical contributions of the flanking contextual sequences, with the downstream more predominant than the upstream sequences, in terms of both strength and length. To further quantify and confirm their contributions, we perturb nearby regions alternately and evaluate their impact on the prediction performance (Figure 6.9d-f and Figure 6.5). We observed that the central M0 (+/- 2bp) and downstream D1 [+3bp, +7bp] regions play the most important role in predicting 6mA across different species, which is in line with cis-regulatory patterns elucidated by SM-CAP. We noticed some patterns are shared between the two eukaryotes, e.g., GAGG [-1bp, +2bp] as shown in Figure 6.9a and 6.9b. The normalized scoring maps thus present a good summary of underlying conventional conserved motifs, which might co-regulate 6mA. Further examination of the salient patterns revealed by SM-CAP shows that these patterns are more discriminative, and account for more than 45% of the 6mA sites with odds ratio of more than 20 for A. thaliana and D. melanogaster, respectively. It is noted that conventional conserved motifs could be a special case of the patterns SM-CAP can capture. An example in point is the well-known motif GATC [-1bp, +2bp] in E. coli [14], as identified successfully by SM-CAP as well.

6.2.7 A Whole-genome 6mA Catalog Made by DeepM6A

Lastly, we use DeepM6A to make genome-wide 6mA prediction by scanning the whole genomes (Supplementary Tables 1-3). To the best of our knowledge, this is the first attempt for the systematic identification of 6mA de novo sites based upon just sequence information. A substantial amount of novel 6mA candidate sites can be mapped. We expect the catalog of these novel potential 6mA sites would be useful for investigating the functional roles of 6mA. For instance, further enrichment analysis using this in silico catalog reveals that 6mA is enriched in tRNA regions of A. thaliana and D. melanogaster (6.10b and d). We experimentally confirm this novel association by identifying a methyltransferase for m6A in tRNA (Gu et al., in preparation). However, using just the in vivo validated 6mA sites reported to date leads to less or no significant enrichment (Figure 6.10a and c).

6.3 Discussion

In conclusion, our novel DeepM6A method can predict 6mA sites with high accuracy and reliability. It is noted that our prediction is purely based on sequence information. In other words, only cis-effect has been captured. This means that what we predict is the candidacy or potential for being a 6mA site. Whether a candidate is 6mA-ed or not will also depend on many other exogenous trans-effects. Nevertheless, this in silico 6mA candidacy map helps to provide a global view of 6mA events. Our method has uncovered many interesting regulatory patterns that are missed by conventional motif analysis and as such is worth further investigation. Taken together, our work helps to elucidate the regulatory mechanisms of 6mA and contributes new insights to the in-depth exploration of their functional effects.

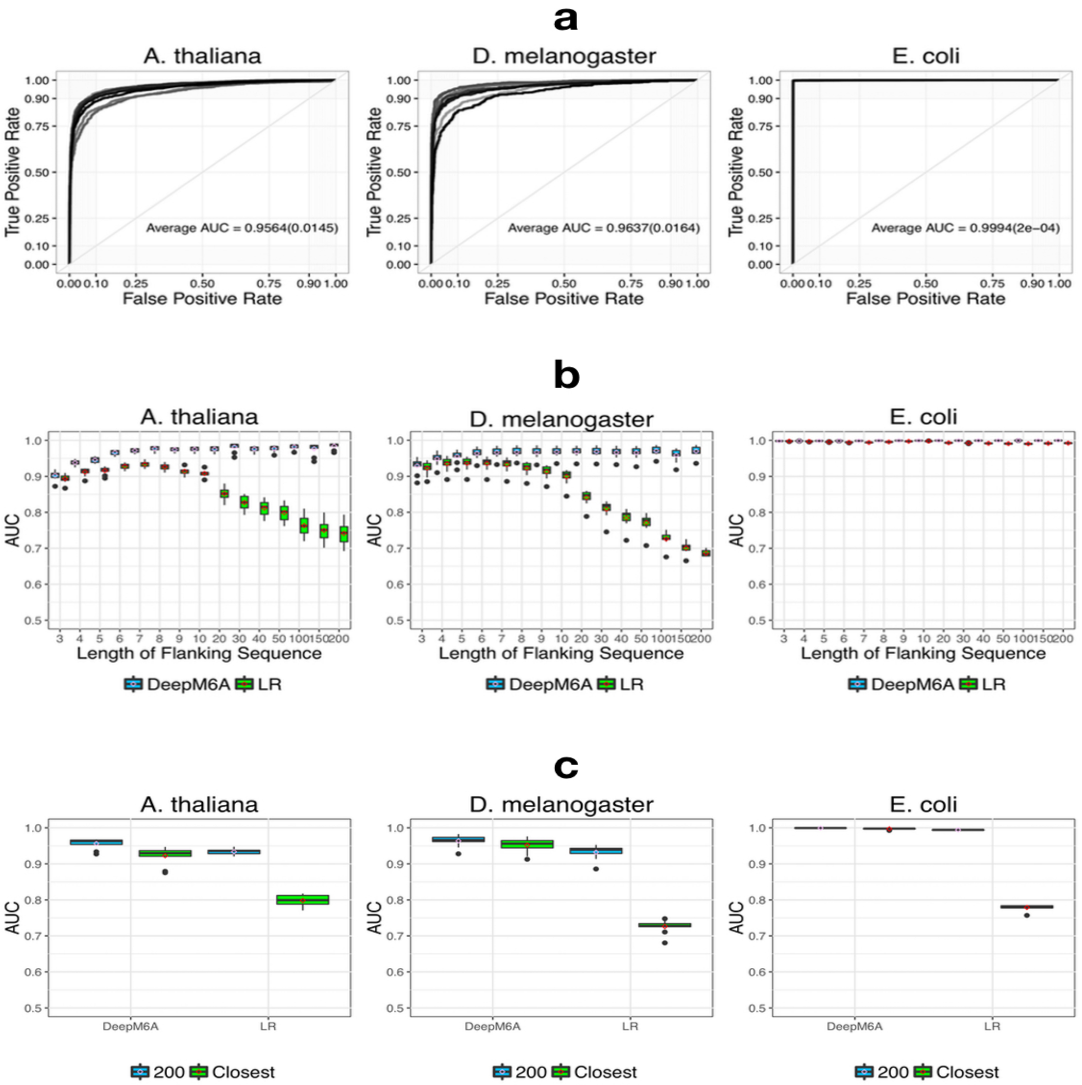


Figure 6.7 The DeepM6A precisely predicts DNA 6mA sites from sequence with single-nucleotide sensitivity for A. thaliana, D. melanogaster and E. coli. a Receiver operating characteristic (ROC) curves for 10 independent experiments and average area under ROC curves (AUC, mean S.D.). b Comparison of DeepM6A versus k-mer based LR across varied contextual sequences. c Impact comparison of remote (200bp) versus closest control cohorts of non-methylated adenines on predictive capacity of DeepM6A and k-mer based LR. Model performance is measured with box plots of AUC over for 10 independent experiments.

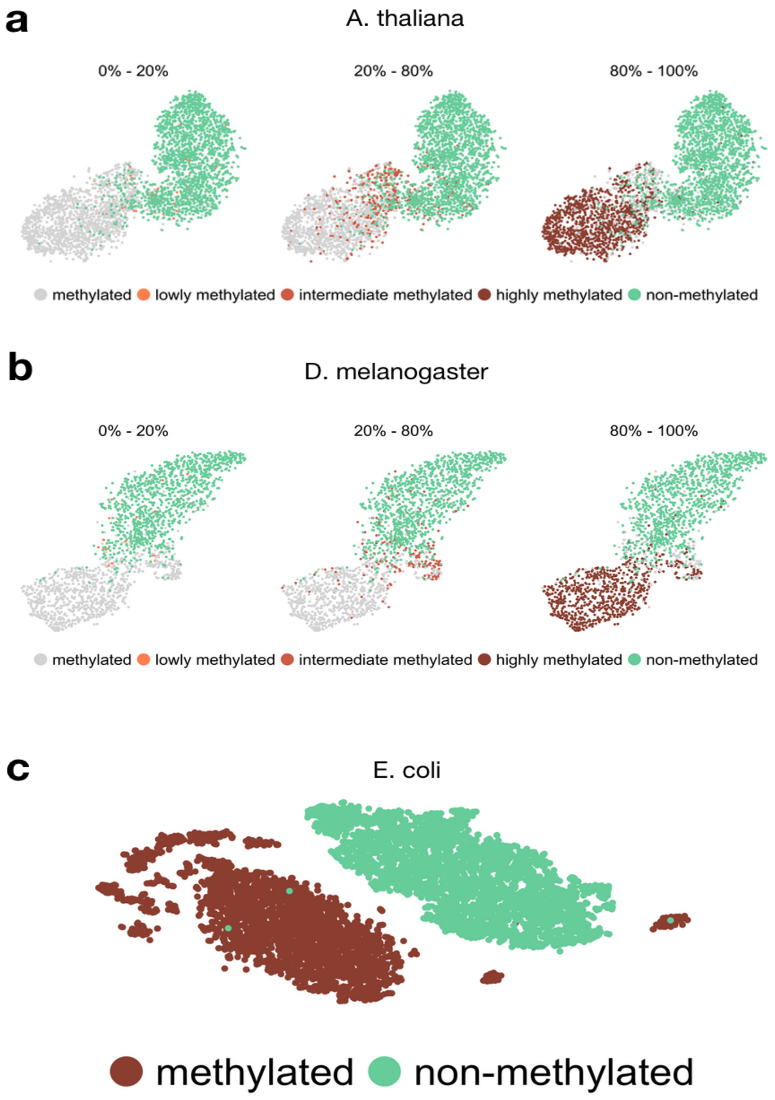


Figure 6.8 t-SNE visualization of the last hidden layer representations of methylations at different levels: lowly (0%-20%), intermediate (20%-80%), and highly (80%-100%). a A. thaliana b D. melanogaster c For E. coli, almost all 6mA sites lie in the highly methylated region.

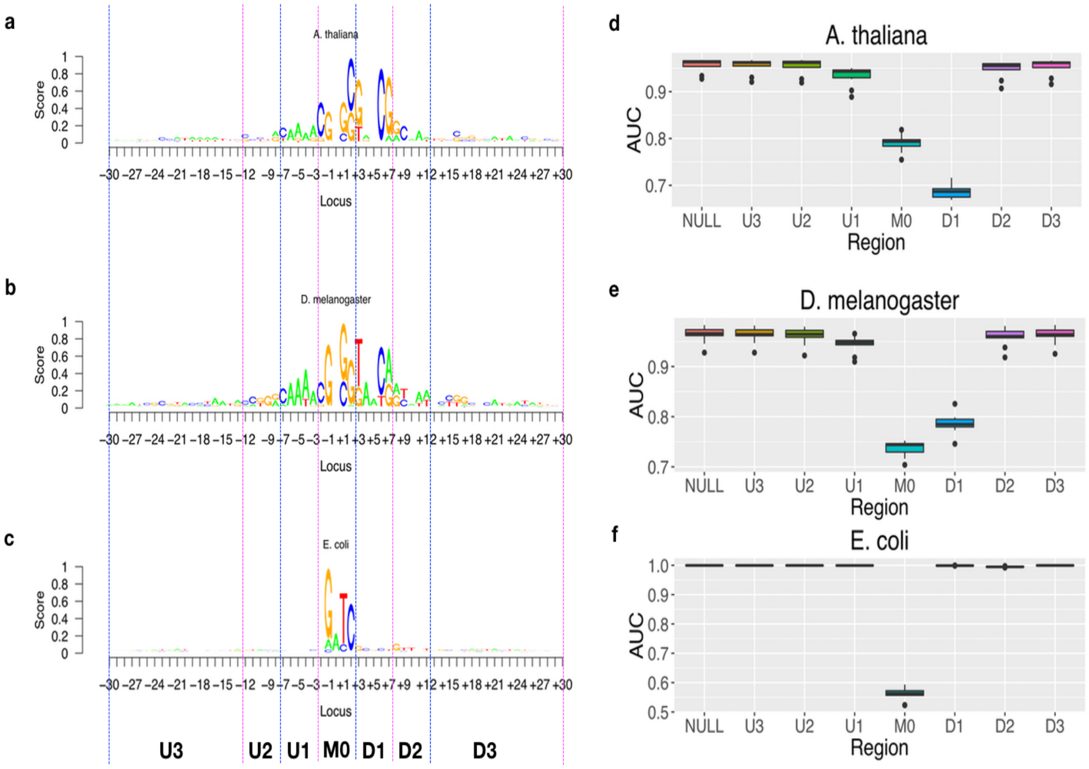


Figure 6.9 DNA motifs and loci revealed by SM-CAP. a-c cis-regulatory patterns (6mA based in 0). d-f Evolution of predictive capacity of DeepM6A when perturbing nearby regions. M0 and D1 are critical regions.

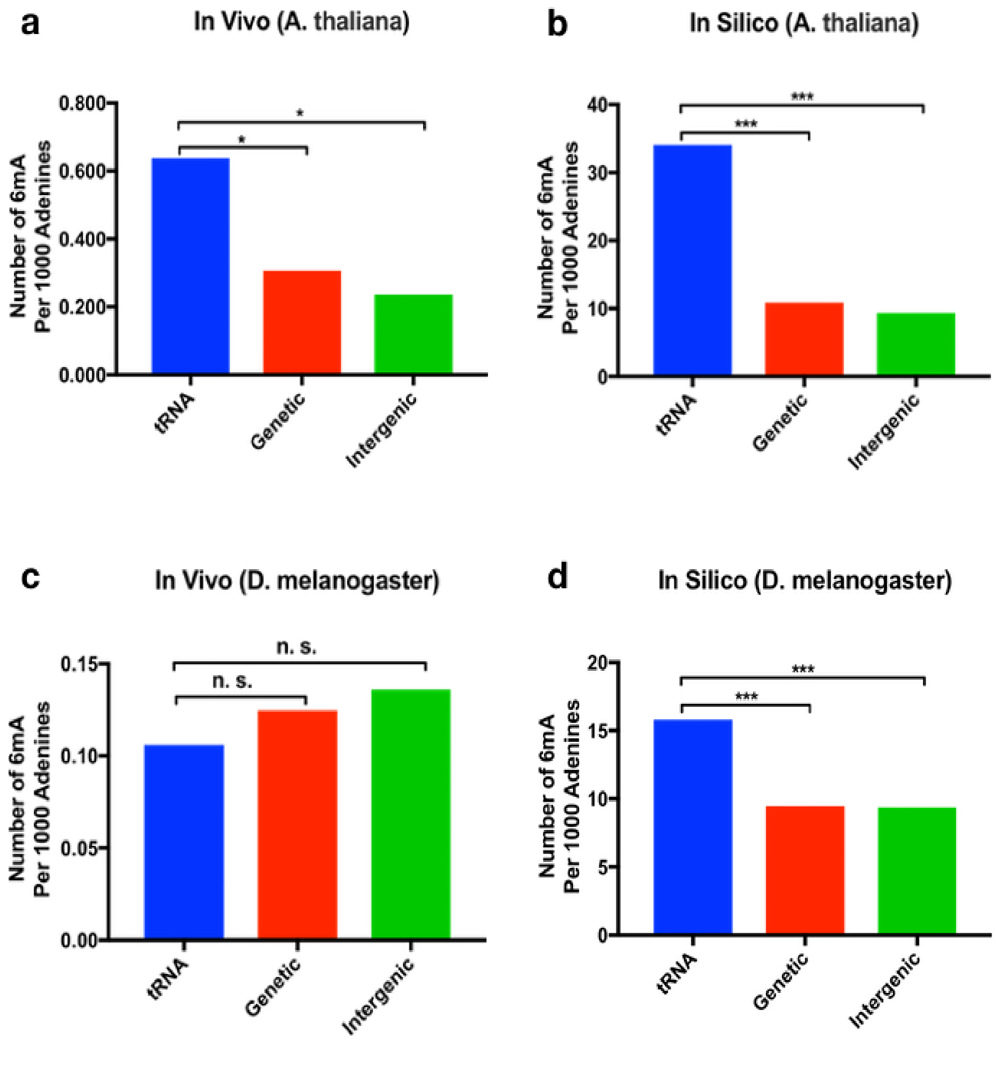


Figure 6.10 Genomic landscape of 6mA. a c In vivo enrichment analysis. b d In silico enrichment analysis. Two-tailed t-test are performed (*$p < 0.05$, ***$p < 0.0005$).

CHAPTER 7

COMPETING RISKS REPRESENTATION IN PEER-TO-PEER LENDING

7.1 Introduction

Online peer-to-peer lending is expected to benefit both investors and borrowers due to their low transaction cost and the elimination of expensive intermediaries. From lenders' perspective, maximizing their return on investment is an ultimate goal during their decision-making procedure. In this dissertation, we explore and address a fundamental problem underlying such a goal: how to represent the two competing risks, charge-off and prepayment, in funded loans. We propose to model both potential risks simultaneously, which remains largely unexplored until now. We first develop a hierarchical grading framework to integrate two risks of loans both qualitatively and quantitatively. Afterwards, we introduce an end-to-end deep learning approach to solve this problem by breaking it down into multiple binary classification sub-problems, which are amenable to both feature representation and risks learning. Particularly, we leverage deep neural networks to jointly solve these sub-tasks, which leads to the in-depth exploration of the interaction involved in these tasks. To our knowledge, this is the first attempt to characterize competing risks for loans in peer-to-peer lending via deep neural networks. The comprehensive experiments on real-world loan data show that our methodology is able to achieve an appealing investment performance by modeling the competition within and between risks explicitly and properly. The feature analysis based on saliency maps provides useful insights into payment dynamics of loans for potential investors intuitively.

7.2 Related Work

Some related studies are discussed as follows.

Peer-to-Peer Lending: In practice, lenders expect that their invested loans get fully funded and issued successfully. Thus, the prediction of fully funded loans was explored [77]. They are able to aid in loan evaluation in terms of investment efficiency. Another important research direction in P2P lending is the risk assessment, which can help lenders reduce the potential risk of investment. The common strategy is to group loans into two categories based on the charge-off risk. Afterwards, large numbers of classifiers are leveraged to conduct classification learning [22, 47, 183]. For example, some easy-to-interpret methods like logistic regression were proposed to model the credits of loans and borrowers from an economic perspective [47, 183]. Byanjankar *et al.* also proposed a credit scoring model based on artificial neural networks to detect potential charged-off loan applications [22]. More recently, Zhao *et al.* considered fully-funded probability, charge-off risk and winning-bid probability simultaneously to propose a multi-objective portfolio optimization approach [182]. These works concentrated on the overall charge-off risk yet ignored its fine-grained survival time in terms of risk modeling. More importantly, the risk of prepayment is in lack of exploration in the above researches.

Competing Risks Analysis:

The popular methods are a cause-specific competing risks model [94] and a proportional hazards model proposed by Fine and Gray [54]. Regarding competing risks of charge-off and pre-payment in loan data, the above methods can be naturally used for learning payment dynamics. Survival analysis, however, doesn't emphasize the competition within and between multiple risks. Our proposed framework can integrate such consideration into the modeling effectively. Sirignano *et al.* proposed to capture the evolution of mortgage statuses by modeling the transition probability using deep learning [146]. Our model would be augmented by the consideration of intermediate state transition context provided that intermediate statuses are available. Unfortunately, in P2P lending, the trajectory of the loan state process is not publicly accessible. The final status or a snapshot status of loans is only available at the released time point. A cause-specific survival model [87] has been utilized to capture competing nature of prepayment and default for mortgage risks by introducing two position intensities [138]. The focus of the above work is on obtaining a better understanding of economic factors associated with mortgage risks via modeling survival time as discussed before [138]. Actually, a cause-specific survival analysis is also used as a baseline in our study where lognormal and exponential intensities are applied. Our preliminary work simply assumed that prepaid loans are better than charged-off ones and considered only closed loans with the single repayment term [156]. In this chapter, we propose the hierarchical grading and further consider censored loans with multiple repayment terms comprehensively.

Ordinal Regression: Ordinal regression tries to solve an intermediate problem between regression and classification [52, 53, 139]. The target variable is ordinal, and the relative order between different values is emphasized in model learning[1]. Li *et al.* proposed a general framework to systematically reduce the ordinal regression to a series of binary classification [95]. It enables well-tuned binary classification approaches to be readily transformed into appealing ordinal regression algorithms with sound theoretical and empirical support. Under this framework, different ordinal regression algorithms like perceptron-based and SVM-based methods [76, 142] are proposed to solve related problems.

Different from classical ordinal regression problem, our work is the first attempt to assess two competing risks of time-to-event loan data hierarchically. Besides, we introduce masking layers to integrate censored data and multi-term loans into the learning of deep neural networks.

7.3 Methodology

In this section, we describe the procedure of competing risks grading in P2P lending and then present the representation learning approach of deep neural networks.

7.3.1 Hierarchical Grading of Competing Risks

Let D represent the original dataset of loans including N loans $\{(\mathbf{x}_i, s_i, t_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X} = \mathbb{R}^d$ is the input feature vector, and $s_i \in \mathcal{S} = \{1, 2, 0\}$ is the status of loans with 1, 2, and 0 being charged-off, fully-paid and censored, respectively. $t_i \in \mathcal{T} = \{0, 1, \dots, T-1, T\}$ is the total received payment count, where T is the official scheduled term of a loan. Charged-off and fully-paid loans are also called closed loans here. In practice, few loans cannot receive any payment, we thus have

[1] https://en.wikipedia.org/wiki/Ordinal_regression

$\mathcal{T} = \{1, \ldots, T-1, T\}$ for simplicity. It is noted that fully paid loans involve two types of payments, i.e., prepayment and scheduled payment as illustrated in Figure 7.3. The prepayment is thought of as that loans are paid off before the official due date, whereas loans paid off strictly based on the schedule are the type of scheduled payment. As with the risk of charge-off, prepayment is another risk existing in investments since less interest can be secured compared with scheduled payment. Besides, the loans without a definite status of charge-off or full-payment are referred to as censored loans. As diagrammed in Figure 7.3, the maximum of total received payment counts for charged-off and censored loans is $T-1$ since T monthly payments are equivalent to the status of scheduled full-payment.

To model both status and payment count of loans, we propose a risk grading rule $g : \mathcal{S} \times \mathcal{T} \to \mathcal{Y}$ as follows:

$$y_i = \begin{cases} d_{t_i}, & s_i = 1 \\ p_{t_i}, & s_i = 2 \\ c_{t_i}, & s_i = 0 \end{cases} \tag{7.1}$$

where d, p and c stand for default (i.e., charge-off), full payment and censor, respectively. s_i and t_i are the status and survival time as mentioned before. Consequently, we have the corresponding converted dataset $\mathcal{O} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $y_i \in \mathcal{Y} = \{c_1, c_2, \ldots, c_{T-1}, d_1, d_2, \ldots, d_{T-1}, p_1, p_2, \ldots, p_T\}$. For closed loans, we have definite final statuses. As per usual, we adopt $\prec$ as the grading relation. For closed loans, risk grades follow $d_1 \prec d_2 \prec \ldots \prec d_{T-2} \prec d_{T-1}$, and $p_1 \prec p_2 \prec \ldots \prec p_{T-1} \prec p_T$. The philosophy behind such a grading strategy is straightforward: in regards to loans of the same status, the more monthly payments are received, the more desirable loans are. Therefore, the identical loans receiving more payment times are assigned higher grades accordingly. For censored loans, their categories are highly related to the observation time point, which leads to the uncertainty of final payment status and payment count. We thus don't place grading on different categories here. In Section 7.3.3, we will revisit this problem for embedding censored loans into the unified representation of closed loans.

7.3.2 Methodology Framework Overview

Figure 7.1 is given to facilitate the understanding of our modeling pipeline. The framework mainly consists of four components: (1) Generation of risk grades based on loan status, survival time and loan term. (2) Conversion from the hierarchical ordinal grades to multiple binary outputs of multiple terms. The developed network architecture is reported in Figure 7.4. (3) Competing risks prediction and evaluation for held-out loan data. (4) Model interpretation with respective to feature importance visualization.

7.3.3 Deep Learning Approach

Mathematically speaking, the risk grading can be defined to search for a mapping rule from input features to the grading category $h(\cdot) : \mathcal{X} \to \mathcal{Y}$ such that

$$\arg\min_{h} \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{y_i,h(x_i)} \tag{7.2}$$

where $\mathcal{C}$ is a defined $K \times K$ cost matrix with $\mathcal{C}_{y_i,h(x_i)}$ that a sample (x_i, y_i) is predicted as category $h(x_i)$. It is reasonably assumed that $\mathcal{C}_{y_i,h(x_i)} = 0$ if $y_i = h(x_i)$, otherwise,

$\mathcal{C}_{y_i,h(x_i)} > 0$ [95]. Naturally, the cost mechanisms should give more penalty to erroneous prediction of the grading category. That is to say, $\mathcal{C}_{y_i,h(x_i)-1} \geq \mathcal{C}_{y_i,h(x_i)}$ if $h(x_i) \leq y_i$ and $\mathcal{C}_{y_i,h(x_i)} \leq \mathcal{C}_{y_i,h(x_i)+1}$ if $h(x_i) \geq y_i$. A popular choice is absolute cost matrix defined by $\mathcal{C}_{y_i,h(x_i)} = |y_i - h(x_i)|$.

For the purpose of competing risks learning and predictive inference, we detail five procedures as follows.

Conversion from Closed Grading Categories to Multiple Binary Outputs

The loan data containing K grading categories can be converted to K binary classification problems for each type of status. Concretely speaking, given hierarchical grading dataset of closed loans $\mathcal{O}_{closed} = \{(x_i, y_i)\}_{i=1}^{N_{closed}} \subset \mathcal{O} = \{(x_i, y_i)\}_{i=1}^{N}$ where N_{closed} is the number of closed loans, the specific training data for binary classifier k is $\mathcal{B}^k = \{(x_i, y_i^k, w_i^k)\}_{i=1}^{N_{closed}}$ where y_i^k indicates whether the category of sample (x_i, y_i) is larger than or equal to category k, and w_i^k is the corresponding weight. Formally, y_i^k is defined as follows:

$$y_i^k = \begin{cases} 1, & y_i \geq k \\ 0, & \text{otherwise} \end{cases} \tag{7.3}$$

To ensure that the original cost of risk grading is bounded by weighted zero-one loss on converted examples [95], it is specified that $w_i^k = |\mathcal{C}_{y_i,h(x_i)} - \mathcal{C}_{y_i,h(x_i)+1}|$. Since absolute cost matrix is adopted here, $w_i^k = 1$. Another equivalent interpretation is that if the category of a sample is $y_i = k$, it is grouped into lower-grading categories $\{1, 2, \ldots, k-1\}$ as well. In this case, the final target vector is $\mathbf{z} = (1, \ldots, 1, 0, 0, 0)$, where z_i $(1 \leq i \leq k)$ is set to be 1 and the remaining elements are zeros. It also implies that the loan has gone through the previous monthly payments. In this manner, the final predicted probability vector is thus expected to share the following property: $\hat{z}_i(i \leq k)$ approaches 1 and $\hat{z}_i(i > k)$ is close to 0. For types of charge-off and full payment, we have $K = T-1$ and T, respectively, as derived by Equation 7.1. We then further propose to fuse multiple binary outputs of two statuses to generate a unified representation (a vector of $2T-1$ elements) with hierarchical grading constraints (e.g., charge-off and pre-payment of the loan with 60-month term as diagrammed in Figure 7.2). In this case, $y_i = d_k$ and $y_i = p_k$ can be represented as $(\mathbf{1}_{d_1 \sim d_k}, \mathbf{0}_{d_{k+1} \sim d_{T-1}}, \mathbf{0}_{p_1 \sim p_T})$ and $(\mathbf{0}_{d_1 \sim d_{T-1}}, \mathbf{1}_{p_1 \sim p_k}, \mathbf{0}_{p_{k+1} \sim p_T},)$, respectively.

Representation of Censored Loans For the dataset of censored loans $\mathcal{O}_{censored} = \{(x_i, y_i)\}_{i=1}^{N_{censored}} \subset \mathcal{O} = \{(x_i, y_i)\}_{i=1}^{N}$ where $N_{censored}$ is the number of censored loans, we embed $y_i = c_k$ $(1 \leq k \leq T-1)$ into the representation scheme of closed loans with $\mathbf{z} = (\mathbf{1}_{d_1 \sim d_k}, \mathbf{0}_{d_{k+1} \sim d_{T-1}}, \mathbf{1}_{p_1 \sim p_k}, \mathbf{0}_{p_{k+1} \sim p_T})$. Here $\mathbf{1}_{d_1 \sim d_k}$ and $\mathbf{1}_{p_1 \sim p_k}$ indicate that the status of charge-off or full-payment is possible and k monthly payments have been received until the observation time point. However, the final payment times remain unknown yet. Thus, $\mathbf{0}_{d_{k+1} \sim d_{T-1}}$ and $\mathbf{0}_{p_{k+1} \sim p_T}$ here lead to a biased representation. To correct this, we further introduce a masking vector $\mathbf{m} = (\mathbf{0}_{d_1 \sim d_k}, \mathbf{1}_{d_{k+1} \sim d_{T-1}}, \mathbf{0}_{p_1 \sim p_k}, \mathbf{1}_{p_{k+1} \sim p_T})$ where $m_i = 1$ if element i is masked, $m_i = 0$ otherwise. It can be utilized to exclude biased elements from the procedure of model learning accordingly (e.g., censored loans of 60-month term as shown in Figure 7.2).

Fusing Multiple Terms with Padding In practice, loans in P2P lending usually involve multiple scheduled terms. For example, two types of 36-month and

60-month are observed in Lending Club. To integrate loans of multiple terms for a unified representation, we propose to perform zero padding on representation vector of short terms. Specifically, given L different scheduled terms, $T_l \in \{T_1, \ldots, T_L\}$, we have the maximum term $T_{max} = max(\{T_1, \ldots, T_L\})$. $y_i = d_k$ for loan i of term T_l can be zero-padded from $(\mathbf{1}_{d_1 \sim d_k}, \mathbf{0}_{d_{k+1} \sim d_{T_l - 1}}, \mathbf{0}_{p_1 \sim p_{T_l}})$ to $(\mathbf{1}_{d_1 \sim d_k}, \mathbf{0}_{d_{k+1} \sim d_{T_l - 1}}, \mathbf{0}_{d_{T_l} \sim d_{T_{max}-1}}, \mathbf{0}_{p_1 \sim p_{T_l}}, \mathbf{0}_{p_{T_l+1} \sim p_{T_{max}}})$. As $\mathbf{0}_{d_{T_l} \sim d_{T_{max}-1}}$ and $\mathbf{0}_{p_{T_l+1} \sim p_{T_{max}}}$ are in no sense, we also introduce a masking vector $(\mathbf{0}_{d_1 \sim d_k}, \mathbf{0}_{d_{k+1} \sim d_{T_l - 1}}, \mathbf{1}_{d_{T_l} \sim d_{T_{max}-1}}, \mathbf{0}_{p_1 \sim p_{T_l}}, \mathbf{1}_{p_{T_l+1} \sim p_{T}}$ in the same manner, as mentioned in last section. Figure 7.2 illustrates this point by loans of 36-month term.

Deep Neural Networks Armed with the above analysis, we leverage deep neural networks to conduct a series of binary classification. As shown in Figure 7.4, the designed networks consist of an input layer of d-dimension input feature nodes, fully connected shared and parallel individual layers with masking outputs, and L parallel output layers of $2T_{max} - 1$ nodes. Shared layers learn common feature representation across multiple terms, whereas individual layers target specific patterns for loans of the corresponding term. Different from traditional multi-class classification neural networks with softmax activation function of the output layer, the sigmoid function is adopted for the underlying architecture. The main idea is to enable output probability of different classifiers to be estimated independently without constraints with each other. The corresponding loss is the widely-used binary cross entropy function for loans of term l:

$$\begin{aligned} \mathcal{J}_l = -\frac{1}{N_l} \sum_{i=1}^{N_l} \sum_{v \in V} (1 - m_{li}^v) \Big[& y_{li}^v \log f_l^v(x_{li}) + \\ & (1 - y_{li}^v) \log(1 - f_l^v(x_{li})) \Big] \end{aligned} \tag{7.4}$$

where $V = \{d_1, \ldots, d_{T_{max}-1}, p_1, \ldots, p_{T_{max}}\}$. The probability of output node $v \in \{d_1, \ldots,$

$d_{T_{max}-1}, p_1, \ldots, p_{T_{max}}\}$ is denoted as $f_l^v(\cdot) \in [0, 1]$. Since deep neural networks with competing risks representation are developed here, we call it crDNN for brevity. Finally, we can obtain the total loss function over different terms:

$$\mathcal{J} = \sum_{l=1}^{L} \mathcal{J}_l \tag{7.5}$$

Predictive Inference Suppose a new loan x_{lj} is given, we aim to estimate three metrics: charge-off or default probability $p(x_{lj})$, multi-class probability $p^v(x_{lj})$ and the survival time $s(x_{lj})$.

The estimation formula of $p(x_{lj})$ can be naturally derived as follows:

$$\mathrm{p}(x_{lj}) = \frac{f_l^{d_1}(x_{lj})}{f_l^{d_1}(x_{lj}) + f_l^{p_1}(x_{lj})} \tag{7.6}$$

When the binary classifiers $f_l^v(\cdot)$ are consistent, i.e., $f_l^{d_1}(x_{lj}) \geq \ldots \geq f_l^{d_{T_l-1}}(x_{lj})$ and $f_l^{p_1}(x_{lj}) \geq \ldots \geq f_l^{p_{T_l}}(x_{lj})$, multi-class probability $p^v(x_{lj})$ can be estimated by

$$p^v(x_{lj}) = \begin{cases} f_l^v(x_{lj}) - f_l^{v+1}(x_{lj}), & v \in V_l, v \notin \{d_{T_l-1}, p_{T_l}\} \\ f_l^v(x_{lj}), & v \in \{d_{T_l-1}, p_{T_l}\} \end{cases} \tag{7.7}$$

where $V_l = \{d_1, \ldots, d_{T_l-1}, p_1, \ldots, p_{T_l}\}$. Normalization on $p^v(x_{lj})$ is introduced to ensure $\sum_{v \in V_l} p^v(x_{lj}) = 1$.

The survival time probability distribution $s^k(x_{lj})$ can be estimated by

$$s^k(x_{lj}) = \begin{cases} f_l^{p_{T_l}}(x_{lj}), & k = T_l \\ f_l^{d_k}(x_{lj}) + f_l^{p_k}(x_{lj}), & k \in \{1, 2, \ldots, T_l - 1\} \end{cases} \tag{7.8}$$

$s(x_{lj})$ can be estimated by

$$s(x_{lj}) = \min\{k : s^k(x_{lj}) < 0.5\} \tag{7.9}$$

The current modeling scheme, however, cannot explicitly ensure the strict consistency among different binary classifiers theoretically. As in [95, 126], we apply the above formula to prediction directly since the consistency can be observed well in practical experiments. Furthermore, The theoretical consistency will bring about the significant modeling complexity.

7.4 Experiment

In this section, we describe our experimental procedure and report empirical evaluation results of the proposed framework on the dataset from Lending Club. The whole evaluation procedure is composed of two aspects: 10-fold cross-validation and moving-time window experiments.

7.4.1 Experimental Data and Preprocessing

We downloaded loan data as of Quarter 4, 2016[2] from Lending Club. There are a total of $1,321,864$ loan records, which involve types of 36-month and 60-month regarding the scheduled term. Large numbers of loans were issued after the year of 2015. Particularly, more than 60% of loans are still in progress (*Current* in Lending Club) among all issued loans, which are often called censored data in classical survival analysis. In addition to censored loans, we also focus on closed loans, which mainly

[2] https://www.lendingclub.com/info/download-data.action

include statuses of charge-off (*Default* and *Charged Off* in Lending Club) and full payment. There are still a tiny proportion of other statuses such as in grace period, late and so on, which are filtered out for simplicity.

Since released datasets are statistics of historical loans, they involve many features unavailable in profiles of loans when they are listed for investment. In order to simulate the real-world investment scenario to the maximum extent, we filter out features of this kind for facilitating the payment dynamics prediction of loans based on the learned model.

Then we group the related features of loans into numerical and categorical clusters. Regarding numerical features (e.g., loan amount, FICO score), we conduct standardization for training data to transform features to have zero mean and unit variance. Such preprocessing is amenable to the acceleration of the optimization procedure. With the aid of original center and scale of training features, we standardize numerical features of both validation and test data accordingly. In regards to categorical features (e.g., grade and purpose), we utilize one-hot encoding for training, validation and testing data to represent different categories for the same feature. Afterwards, we fuse numerical and categorical features together and generate the final input feature set.

After data cleaning, we finally have a total of $1,269,019$ loans for study as detailed in Table 7.1. To tune hyper-parameters for the proposed framework and proceed with the modeling evaluation, we conduct stratified 10-fold cross validation. What's more, we split the whole dataset with moving time cutoffs as detailed in Table 7.2.

Table 7.1 Statistics of Loans of Interest

	Charge off	Full Payment	Censor	Total
36-month	58,085 (6.39%)	312,158 (34.33%)	538,981 (59.28%)	909,224
60-month	37,096 (10.31%)	78,642 (21.86%)	244,057 (67.83%)	359,795
Total	95,181 (7.5%)	390,800 (30.8%)	783,038 (61.7%)	1,269,019

Table 7.2 Statistics of Loans Issued after the Cutoff Time Point

	07/01/14	09/01/14	11/01/14	01/01/15	03/01/15	05/01/15
Charge off	48,716	42,643	36,749	33,086	27,166	21,630
Full Payment	163,600	144,211	126,300	114,711	97,276	82,095
Censor	731,919	710,652	686,729	667,892	634,529	596,853
Total	944,235	897,506	849,778	815,689	758,971	700,578

7.4.2 Experimental Results

Evaluation Metrics We consider three-fold comparisons: (a) Regarding the binary classification of charged-off and fully-paid loans, the classical area under receiving operating curves (AUC@ROC) [151, 154, 157] and precision-recall curves (AUC@PR) [39] are adopted.

(b) The concordance index (C-index), which quantifies the quality of rankings with censored data being considered, is a standard performance measure for model assessment in survival analysis [72]. (c) In Subsection 7.4.2, the return on investment (ROI) are analyzed comparatively in terms of the proper modeling of competing risks.

Baseline Algorithms We compare our proposed framework with the following schemes: (1) Lending Club (LC) has its own grading and evaluation system. Each issued loan is assigned a grade from A to G (five sub-grades per grade) with a matched interest rate. Generally speaking, the higher the interest rate is, the riskier the corresponding loan is. (2) Logistic regression (LR) is frequently utilized for risk evaluation in general purpose credit scoring and P2P lending study [47,183]. In order to facilitate the training with the large-scale dataset, we construct a simple neural network with one input layer and sigmoid activation function with GPU acceleration. Stochastic gradient descent [19] with Nesterov momentum of 0.9, learning rate of 0.01 and decay rate of 10^{-6} and modern anti-overfitting technique dropout [150] are adopted. (3) To explore the role of hierarchical grading regularization, we also design a multi-class deep neural network without grading constraints (mcDNN), which has the same architecture and hyper-parameter settings (detailed in next section) with crDNN. The multinomial cross-entropy is adopted as loss function for model learning. The activation function of the output layer is the softmax function. (4) Competing risks based survival analysis (CRSA) [8,94] is recently applied to credit scoring [111, 177]. There are cause-specific model [87,138] and alternative proportional hazards model proposed by Fine and Gray [54]. The lognormal and exponential distributions are observed for risks of charge-off and pre-payment, respectively. In observance of specific well-studied probability distributions, the parametric modeling of risks (e.g., cause-specific) is more powerful to capture the payment dynamics than non-parametric or semi-parametric (e.g., proportional hazards model) modeling strategies are. Thus, we adopt the former cause-specific model. The basic R routine package

‘CFC’ [110] is employed to implement the learning procedure. The maximum number of subdivisions $N_{max} = 400$ and the threshold for relative integration error $rel.tol = 1e - 04$.

Experimental Setting and Hyper-Parameter Tuning We utilize python libraries TensorFlow 1.0.0 and Keras 2.0.9 to build the architecture of deep neural networks. NVIDIA Tesla K80 GPU with the memory of 12GB is used for model training. There are a total of 4 fully connected hidden shared layers with 200 nodes in each one. Leaky rectified linear unit with gradient $\alpha = 0.001$ for negative inputs [108] is adopted as activation function of hidden layers. The dropout rates of 4 hidden layers are 0.5, 0.5, 0.4, 0.4, respectively. Individual layers share the same architecture across different terms, which are composed of 2 layers with 200 nodes. The batch size is 128. The maximum of epochs is 500 with early stopping of patience of 50 epochs.

Table 7.3 Overall Performance for the Proposed Methodology and Baselines: Mean (Standard Deviation)

Model	AUC@ROC	AUC@PR	C-index
LC	0.6784 (0.0036)	0.3212 (0.0029)	–
LR	0.7152 (0.0031)	0.3731 (0.0038)	–
mcDNN	0.7088 (0.0028)	0.3516 (0.0029)	0.5101 (0.0022)
CRSA	0.6930 (0.0044)	0.2904 (0.0063)	0.5638 (0.0015)
crDNN	**0.7255** (0.0032)	**0.3914** (0.0040)	**0.5797** (0.0021)

Table 7.4 Matched One-tailed T-test If the Mean of Metrics for crDNN is Greater than Baselines

	AUC@ROC	AUC@PR	C-index
LC	2.20E-16	2.20E-16	–
LR	4.30E-07	2.00E-09	–
mcDNN	2.00E-10	5.40E-15	7.90E-24
CRSA	8.13E-13	2.20E-16	3.94E-13

Performance and Analysis We randomly split the whole dataset into 10 partitions and then do 10-fold cross-validation experiments. This data splitting strategy can at least help to evaluate the extent to which the model captures the underlying payment patterns behind historical loan data. For closed loans, the comparison results are shown in Tables 7.3 and 7.4. Overall, our model can present a sound performance gain on baseline algorithms. For AUC@ROC, crDNN seems to be slightly better than others in terms of discriminating charged-off loans from fully-paid ones. In this case, the comparison on AUC@PR is also reported, which is more amenable to the modeling evaluation for class imbalance [39]. The observed AUC@PR of around 0.4 for the proposed method is appealing given the class imbalance ratio of 1:4 between charge-off and full payment. It is demonstrated that such superiority of the proposed model is more evident against baseline methods. Either AUC@ROC or AUC@PR, however, focuses on the binary class decision, which cannot assess the fine-grained payment dynamics. Therefore, C-index is further introduced to quantify the ranking quality of the proposed method. It is noted that neither LR nor LC is capable of rendering the time-to-event prediction. Thus, we mainly compare crDNN with its variant mcDNN and CRSA. The tremendous dominance of crDNN over mcDNN can be observed when the hierarchical ordinal regularization is taken into account. The evident superiority still holds true for the comparison with CRSA.

The difference in performance of the proposed method and baselines is also evidenced by the statistical significance test as detailed in Table 7.4. Altogether, the proposed method is able to discriminate different loan statuses and predict the time-to-event of loans more accurately as compared with baselines.

Return on Investment under Naive Selection Strategy Apart from the aforementioned comparisons purely based on the risk probability, we further perform in-depth exploration in ROI for this sub-section. ROI is of high concern to investors, which is closely correlated with competing risks in funded loans.

Equated monthly installment (EMI) is commonly adopted as a payment scheme for P2P lending[3]. Formally, the monthly installment is given by $A = P\frac{r(1+r)^T}{(1+r)^T-1}$ where P is the principal (funded amount in Lending Club) and r is the monthly interest rate. T is the total number of monthly installments, which is also the scheduled term of loans. The annual interest rate in this dataset should be transformed to monthly interest rate by being divided by 12. Essentially, a larger proportion of each payment is set aside for interest at the beginning of the amortization schedule compared with the end of the schedule. For ease of notation, we omit some subscriptions but retain unambiguity in the following formulas.

[3] http://www.investopedia.com/terms/e/equated_monthly_installment.asp

The monthly paid interest I_k can be calculated mathematically as

$$I_k = \begin{cases} P \times r, & k = 1 \\ \big[P(1+r)^{k-1} - A\sum_{j=0}^{k-2}(1+r)^j\big]r, & 2 \leq k \leq T \end{cases} \tag{7.10}$$

Now given a loan x, the probability of risk category v to which it belongs is $p^v(x)$ as derived in Equation 7.7, where $v \in \{d_1, \ldots, d_{T-1}, p_1, \ldots, p_T\}$.

Afterwards, the estimated ROI for loan x can be formulated as follows[4]:

$$\widehat{ROI} = \tfrac{1}{P}\Big\{A\sum_{k=1}^{T-1} p^{d_k}(x) + \sum_{k=1}^{T} p^{p_k}(x)\big[P + I_k\big] - P\Big\} \tag{7.11}$$

Equation (7.11) can also be applied to estimate the expected ROI provided by mcDNN and CRSA. Regarding mcDNN, $p^v(x)$ can be replaced by output probability of the corresponding class. For CRSA, the cumulative incidence of hazard can be converted to monthly hazard rate since it is discrete. To be specific, we have charge-off and pre-payment hazard rates $r_{t1}(x)$ and $r_{t2}(x)$ for loan x on monthly payment t. Since they are conditional probabilities, the corresponding $p^v(x)$ for CRSA can be derived as follows:

$$p^v(x) = \begin{cases} r_{11}(x), & v = d_1 \\ r_{12}(x), & v = p_1 \\ r_{k1}(x)\prod_{t=1}^{k-1}\big[1 - r_{t1}(x) - r_{t2}(x)\big], & v = d_k \\ r_{k2}(x)\prod_{t=1}^{k-1}\big[1 - r_{t1}(x) - r_{t2}(x)\big], & v = p_k \\ r_{T2}(x)\prod_{t=1}^{T-1}\big[1 - r_{t1}(x) - r_{t2}(x)\big], & v = p_T \end{cases} \tag{7.12}$$

where $k \in \{1, \ldots, T-1\}$.

[4] Standard net present value (NPV) calculation approach is not adopted here as no records of cash flow are available in the dataset.

With the help of Equation (7.11), we utilize a naive selection strategy to compare the ground-truth ROI of selected loans based on crDNN against that of mcDNN and CRSA. In regards to the grading system of Lending Club and logistic regression, we make use of the estimated probability of charge-off to select loans since they are unable to provide the corresponding categorical probability for the estimation of ROI. Basically, the naive strategy is to set up a parameter *topRate* to control the number of selected loans, and then choose loans from all candidates based on given scores ($\widehat{ROI}$ or full-payment probability).

We conduct monthly selection on loans and monitor two metrics: 1) The proportion of months with monthly ROI of selected loans being larger than that of all loans on the same months. The monthly ROI on calendar month i is defined by Equation (7.13). We call months of this kind *good* months for short. 2) The aggregated ROI across different calendar months is also given by Equation (7.13) accordingly.

$$\begin{aligned}
\mathrm{ROI_i} &= \frac{\sum_{j=1}^{R_i} \Omega_{ij} - \Psi_{ij}}{\sum_{j=1}^{R_i} \Psi_{ij}} \\
\mathrm{ROI} &= \frac{\sum_{i=1}^{M} \sum_{j=1}^{R_i} \Omega_{ij} - \Psi_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{R_i} \Psi_{ij}} \\
R_i &= N_i \times \mathrm{topRate}
\end{aligned} \tag{7.13}$$

where Ω_{ij} and Ψ_{ij} are the corresponding total received payments and funded amount of loan j on month i, respectively. R_i and N_i are the number of selected loans and all loans on month i, respectively.

As shown in Figure 7.5, mcDNN, CRSA and crDNN perform better than both LR and LC over different topRates. Such disparity mainly results from the involvement of more fine-grained category of loans in risks modeling. Multifaceted modeling of competing risks is thus more amenable to estimating the real-world ROIs. Particularly, crDNN achieves the highest proportion of good months and the aggregate ROI over different topRates. The superiority of the proposed method against CRSA and mcDNN justifies the necessity of the simultaneous and explicit modeling of both the competition among multiple risks and underlying ordinal constraints. It's noted that CRSA has a better performance than mcDNN in terms of C-index in Table 7.3, whereas this relationship is reversed in Figure 7.5. This is because survival analysis focuses more on the estimate of survival time, whereas mcDNN doesn't impose any grading regularization effects on different fine-grained risk categories from the survival time end. However, mcDNN is able to model the inherent competition among fine-grained risk categories explicitly as mentioned in Section 7.1. To be concrete, the gap between crDNN and CRSA demonstrates that the explicit modeling of the competition within and between multiple risks is beneficial for improving ROI. The disparate performance between crDNN and mcDNN indicates that the internal grading constraint (time-to-event) involved in competing risks modeling coordinates the learning procedure of deep neural networks better. Furthermore, the ROI-topRate curves seem to be relatively steady across different topRates for LC and LR, whereas they exhibit the decreasing trend with the growth of topRate for other three methods. The proportion of good months for LC and LR is around 0.5 over different topRates. This indicates that the selection of loans based on the lowest interest rate or charge-off probability is not a good solution. That is to say, they have the limited guidance towards loan selection in terms of ROI.

Table 7.5 Performance Comparison: AUC@PR (C-index) for Moving-time Window

Type	07/01/14	09/01/14	11/01/14	01/01/15	03/01/15	05/01/15
LC	0.3661 (-)	0.3647 (-)	0.3593 (-)	0.3565 (-)	0.3477 (-)	0.3353 (-)
LR	0.3896 (-)	0.3926 (-)	0.3846 (-)	0.3837 (-)	0.3849 (-)	0.3728 (-)
mcDNN	0.3651 (0.5187)	0.3636 (0.5128)	0.3577 (0.5116)	0.3549 (0.5081)	0.3493 (0.5129)	0.3418 (0.5094)
CRSA	0.3343 (0.5739)	0.3368 (0.5739)	0.3362 (0.5736)	0.3381 (0.5731)	0.3402 (0.5737)	0.3333 (0.5721)
crDNN	**0.3918 (0.5882)**	**0.3935 (0.5834)**	**0.3932 (0.5927)**	**0.3919 (0.5948)**	**0.3853 (0.5998)**	**0.3789 (0.6028)**

Moving-time Window In addition to the overall 10-fold cross-validation, we proceed to perform evaluation with a moving-time window of test data. Such a preferable test scenario can simulate the real-world situations better. To be specific, loans issued before the cutoff date are used for model development, while those remaining loans are grouped into test dataset for empirical evaluation. The distribution of the number of issued loans, however, is highly skewed to recent years. The high concentration of loans towards the end of time period will lead to unevenly distribution of survival time for those closed loans compared with the overall real-world distribution. In this case, ROI analysis based on survival time of only closed loans will introduce heavy bias into the model evaluation. Thus, we utilize AUC@PR to evaluate the overall status of closed loans and C-Index to evaluate the survival time of all loans including censored loans. More discussions about evaluation metrics are detailed in Section 7.5.

The cutoff dates and the statistics of associated loans are reported in Table 7.2. The overall empirical evaluation results for different time windows are provided in

Table 7.5. In practice, Lending Club issues new loans regularly four times per day[5], whereas the issued dates of loans in our dataset are provided monthly. Additionally, the time limit of issued loans for being fully funded is one month. Thus, we conduct more in-depth comparisons on the monthly basis. Such kind of comparison is designed to simulate the real-world issue of loans even though there is still a slight difference. As C-index can incorporate all loans including censored loans, which is an unbiased metric. Thus, we mainly apply C-index to different issued months for fine-grained evaluation as shown in Figure 7.6. Both overall and fine-grained results show that our method is able to outperform baselines.

Therefore, the proposed method is practically appealing based on empirical repeated cross-validation and moving-time window experiments.

Table 7.6 The Most Salient Features Contributing to the Full-payment of Loans

Feature Type	Feature Name
Numerical	annual_inc, total_acc, fico_range_high
Categorical	sub_grade (A4, A5, B4, A3, B5), addr_state (CO, DC, NH, OR, MT, UT), purpose (wedding, credit card, debt consolidation)

Saliency Maps based Feature Analysis After capturing competing risks, the next step typically is to explore how input features shape the payment dynamics of peer-to-peer loans. Modern saliency maps were proposed originally for visualizing

[5] http://blog.lendingclub.com/investor-updates-and-enhancements

Table 7.7 The Most Salient Features Contributing to the Charge-off of Loans

Feature Type	Feature Name
Numerical	dti, int_rate, installment
Categorical	grade (D, E, F), purpose (small business, medical, moving, other, vacation), sub_grade (G3, E5, D5, E4, D2), addr_state (MS, NY, LA, OK, IN, NE), home_ownership (RENT, OWN)

the way how deep convolutional neural networks can be queried regarding the spatial support of a particular class given a specific image [145]. In this study, we extend saliency maps [145] from convolutional neural networks to the proposed learning architecture for quantifying the contribution of each loan feature in the modeling context of all other features. For a loan with feature vector x_0 and a risk category of interest v, the main task is to figure out how elements of x_0 shape output probability $f_l^v(x_0)$ of category v. For the proposed learning method, the score $f_l^v(x)$ is a highly non-linear function of input x. $f_l^v(x)$, however, can be approximated by a linear function in the closeness of x_0 based on the first-order Taylor expansion $f_l^v(x) \approx w^T(x-x_0)+f_l^v(x_0)$ where w is the first-order derivative of $f_l^v(x)$ with respect to the feature vector x at x_0 as $w = \frac{\partial f_l^v(x)}{x}\Big|_{x_0}$. We call w *saliency value*, which has two points to deliver: 1) the magnitude of the derivative indicates the extent to which the change of the most influential elements of feature vector on probability of output node v; 2) the direction of each element of the derivative shows whether such a change improves or degrades the probability of output node v.

In this sub-section, we pick output $v = p_1$ for the interpretation study. As per our proposed representation scheme and learning architecture, the output probability is corresponding to whether the given loan is fully paid or charged off. We compute the saliency maps for all loans and then average them. It is observed that different features contribute to the category of p_1 in different manners. To be specific, some features of loans are highly correlated to the type of full payment, whereas others are closely related to the type of charge-off as shown in Tables 7.6 and 7.7. For instance, a few salient features involve annual income, sub-grade, grade and purpose. It is found that the higher borrowers' annual incomes are, the more they are prone to pay off their loans. Regarding the grade and sub-grades reported by Lending Club, grades A, B and some of their sub-grades contribute to the full payment of loans, while grades D, E, F lead to charged-off loans to some extent. In addition, different borrowing purposes also exhibit different payment preference. Particularly, loans with the purpose of small business and medical issues might be charged off with higher possibility as opposed to those loans with the wedding, debit consolidation and the payment of credit cards.

7.5 Discussion

In this chapter, we model the competing risks of time-to-event loans in P2P lending. The proposed framework of their hierarchical ordinal grading takes into account both the loan status and the time of event. Then deep neural networks are leveraged as the vital classification engine to train the overall framework. Essentially, multi-class deep neural networks can be also viewed as a simple version of our method without prior ordinal constraints. Such regularizing effects are largely in line with ROI and work well for model learning in practice if fixed payment plans EMI is applied as discussed in Section 7.4.2. It is worthy noting that the correlation between the survival time based grading categories and ROI might be degraded when variable payment plans are adopted. Concretely, when a borrower is able to pay higher payment amounts at his/her discretion over the course of payment periods, the investment performance of loans cannot be arranged on their survival time strictly. The variable payment scheme remains the topic of future research.

In addition, the intermediate state transition has been recently incorporated to capture the evolution of mortgage risks [146]. Our framework doesn't take it into account due to two concerns, i.e., the focus of this study and data availability. The final goal of this chapter is to prioritize listed loans prior to their issue for investors. The intermediate statuses after the issued date cannot be utilized in this scenario. Lack of trajectory of loan state process in the released historical loans also prevents us from modeling transition probability among different statuses. The concept of status transition probability can be used in the trading of notes in the secondary market of P2P platform. This is a very promising research topic to explore in the future, which might provide the up-to-date trading suggestions for investors over the course of payment periods [152].

The proposed framework can be thought of as an alternative approach to modeling competing risks and incorporating explicit competition among different categories as compared with classical survival analysis. Hopefully, It aids in unleashing the power of machine learning algorithms for modeling time-to-event loan data beyond online P2P lending. Concretely, the proposed methodology is rather flexible and other classifiers like denoising neural networks [163], support vector machine [35], and ensemble learning [181, 187] can be applied naturally.

In Subsection 7.4.2, we adopt a naive strategy to select loans and compare their corresponding ROIs. It demonstrates the appealing investment performance of crDNN as compared with other methods. In particular, the superior results for crDNN over mcDNN can be observed. In the scenario of moving-time window experiments, the widely used metric towards accrued return with progressively censored loans being included is the net annualized return (NAR) or its variant[6], which can quantify the performance unbiasedly. As we have no access to actual payment trajectory in the dataset, NAR cannot be calculated here. Fortunately, our internal real-time investment test based on the preliminary version of the proposed method delivers an appealing NAR performance indicated by adjusted NAR of Lending Club so far as it goes. Besides, the portfolio optimization or selection is an important research direction, which also has received certain attention in P2P lending study [182]. However, only charge-off is considered in the risk assessment. Our work can be leveraged for the further study of portfolio optimization in online lending [151].

[6] https://www.lendingclub.com/public/lendersPerformanceHelpPop.action

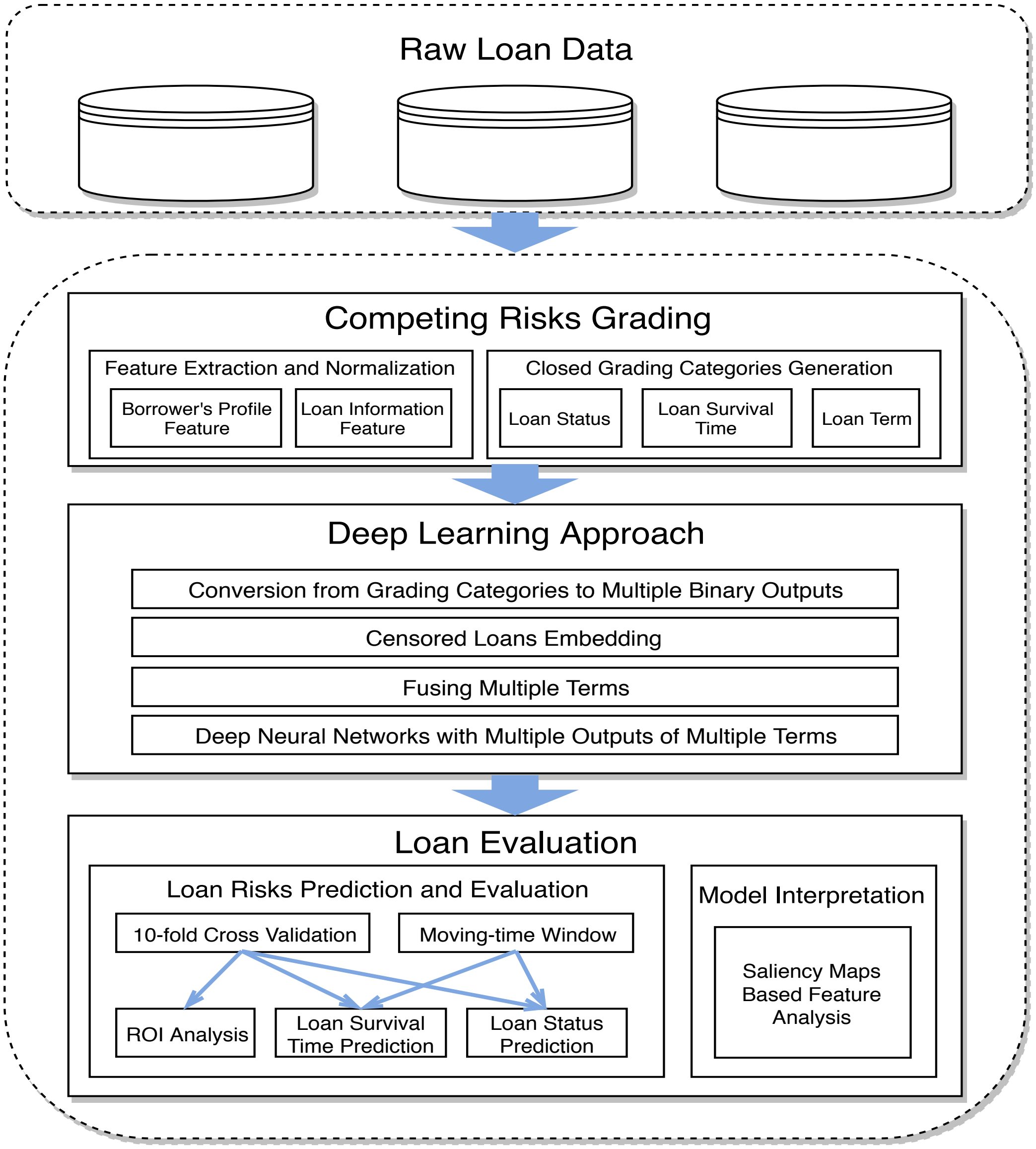

Raw Loan Data
Competing Risks Grading
Feature Extraction and Normalization
Borrower's Profile Feature
Loan Information Feature
Closed Grading Categories Generation
Loan Status
Loan Survival Time
Loan Term
Deep Learning Approach
Conversion from Grading Categories to Multiple Binary Outputs
Censored Loans Embedding
Fusing Multiple Terms
Deep Neural Networks with Multiple Outputs of Multiple Terms
Loan Evaluation
Loan Risks Prediction and Evaluation
10-fold Cross Validation
Moving-time Window
ROI Analysis
Loan Survival Time Prediction
Loan Status Prediction
Model Interpretation
Saliency Maps Based Feature Analysis

Figure 7.1 Framework overview of the proposed methodology.

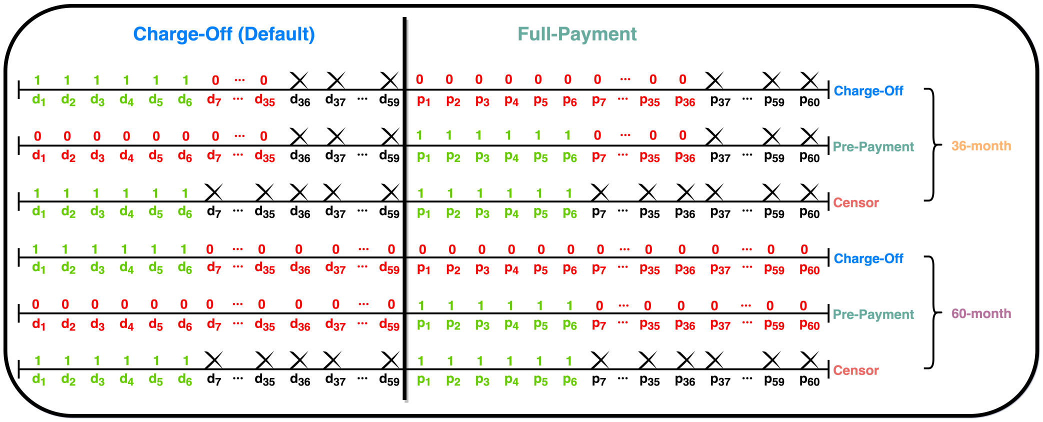

Charge-Off (Default)
Full-Payment
Charge-Off
Pre-Payment
Censor
36-month
Charge-Off
Pre-Payment
Censor
60-month
d_1 d_2 d_3 d_4 d_5 d_6 d_7 ··· d_{35} d_{36} d_{37} ··· d_{59}
p_1 p_2 p_3 p_4 p_5 p_6 p_7 ··· p_{35} p_{36} p_{37} ··· p_{59} p_{60}

Figure 7.2 Illustrative overview of the final representation of charge-off, full payment, and censor for loans with 36-month and 60-month terms. Cross (×) indicates the masking position.

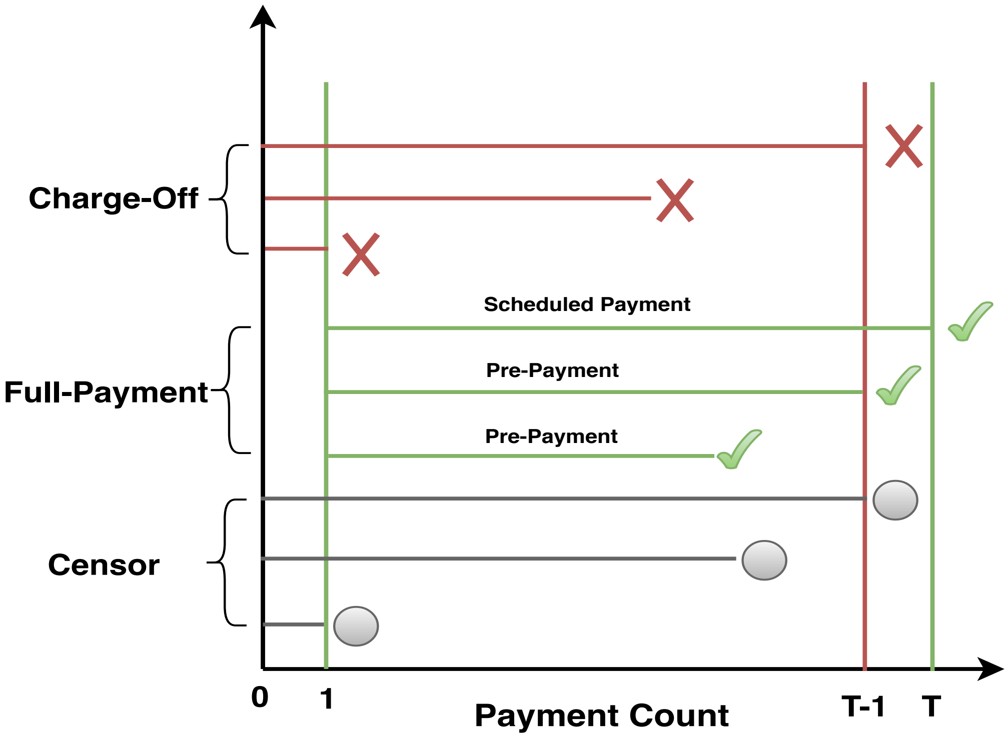


Figure 7.3 Schematic overview of peer-to-peer loan statuses including charge-off, full payment, and censor.

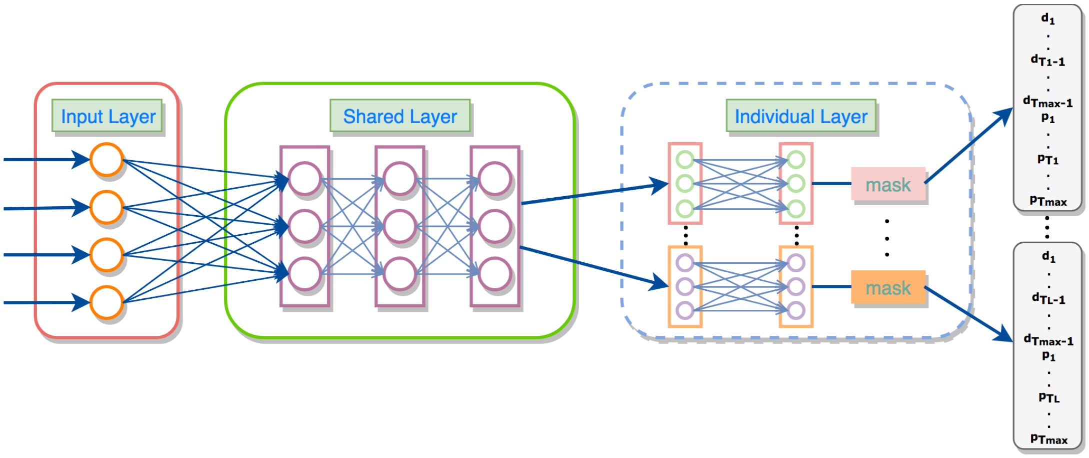


Figure 7.4 The schematic architecture of deep neural networks.

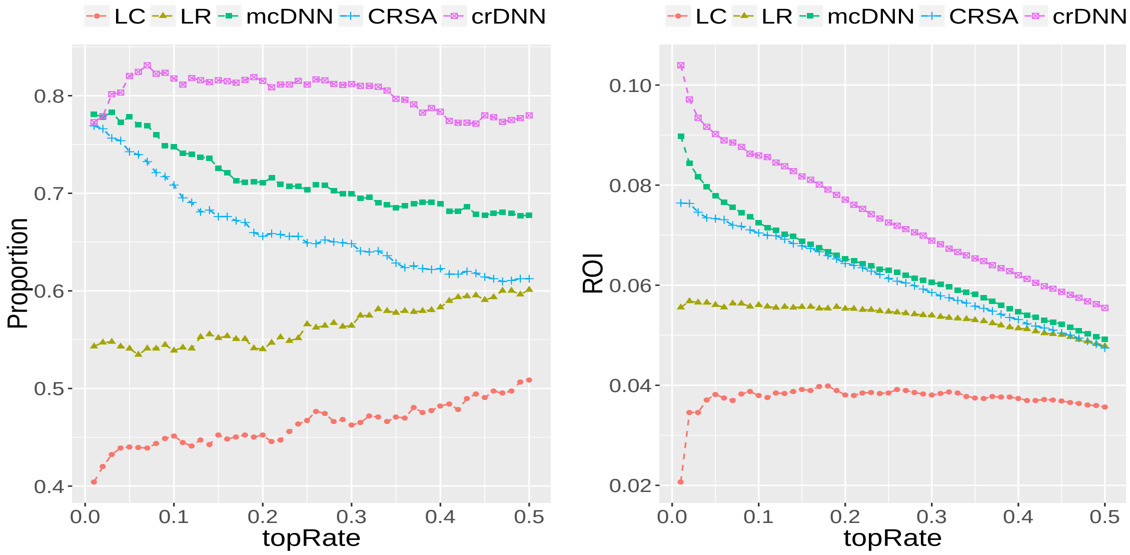


Figure 7.5 Comparisons on the good month and the aggregate ROI for different methods. Each point is averaged over 10-fold cross-validation results.

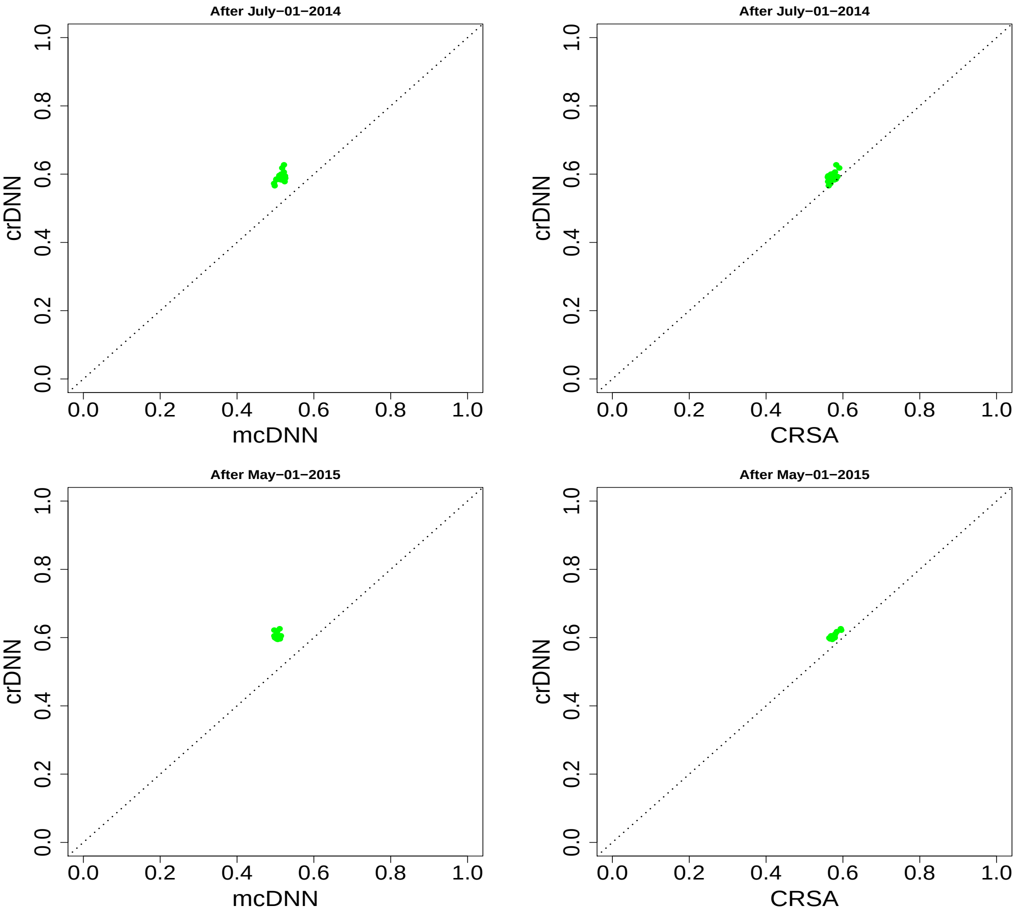


Figure 7.6 The C-index comparisons between crDNN and baselines for loans issued on same months. Each point indicates a monthly comparison.

CHAPTER 8

CONCLUSIONS

This dissertation is dedicated to developing statistical machine learning methodologies for the representation of three genres of spatial and temporal mechanisms with applications to divergent domains. The principal contributions are highlighted as follows.

First of all, in sample-wise dependency, a novel and feasible geographical latent hierarchical probabilistic architecture is proposed to integrate heterogeneous data sources including numerical price, textual comments and geographical location for school district identification. The proposed discrete Markov random field is of vital importance to the performance gain. We also develop a time-aware latent hierarchical model to infer both external and internal components of housing price. The in-depth hierarchical feature ablation analysis is performed to elucidate the critical role of external features in shaping the housing price. For repeated observations, we introduce an item-specific aware model to characterize the involved effects. It works well especially for scenarios with sparse features.

Second, for the feature-wise spatiotemporal interaction, a blended learning scheme is developed to capture the evolution of user activities for intended actions forecasting. A customized convolutional neural networks is also developed to model spatial interactions among neighboring nucleotides for DNA methylation prediction. In addition, a saliency maps based visualization strategy is introduced. It helps to discover retention factors and users' behavior periodicity as well as to visualize regulatory patterns.

Lastly, we propose a framework to represent inherent competition within and between competing risks in time-to-event loan data by modeling the underlying temporal constraints. The work illustrates the general procedure of constructing classifiers for time-to-event loan data, which opens a door towards an alternative to modeling competing risks involved in many problems of practical relevance.

BIBLIOGRAPHY

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation*, pages 265–283, 2016.

[2] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[3] Deepak Agarwal and Bee-Chung Chen. Regression-based latent factor models. In *Proceedings of the 15th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 19–28. ACM, 2009.

[4] Alan G Ahearne, John Ammer, Brian M Doyle, Linda S Kole, and Robert F Martin. House prices and monetary policy: A cross-country study. *International Finance Discussion Papers*, 841, 2005.

[5] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015.

[6] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys*, 51(4):83, 2018.

[7] Wai-Ho Au, Keith Chun Chung Chan, and Xin Yao. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation*, 7(6):532–545, 2003.

[8] Peter C Austin, Douglas S Lee, and Jason P Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016.

[9] Adriano Azevedo-Filho and Ross D Shachter. Laplace's method approximations for probabilistic inferencein belief networks with continuous variables. In *Proceedings of the 10th International Conference on Uncertainty in Artificial Intelligence*, pages 28–36. Morgan Kaufmann Publishers Inc., 1994.

[10] Martin J Bailey, Richard F Muth, and Hugh O Nourse. A regression method for real estate price index construction. *Journal of the American Statistical Association*, 58(304):933–942, 1963.

[11] Timothy L Bailey and Charles Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21(1):51–80, 1995.

[12] Shumeet Baluja, Rohan Seth, D Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th International Conference on World Wide Web*, pages 895–904. ACM, 2008.

[13] Ramesh Baral and Tao Li. Exploiting the roles of aspects in personalized poi recommender systems. *Data Mining and Knowledge Discovery*, pages 1–24, 2017.

[14] Frederic Barras and Martin G Marinus. The great gatc: Dna methylation in e. coli. *Trends in Genetics*, 5:139–143, 1989.

[15] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and Trends® in Machine Learning*, 2(1):1–127, 2009.

[16] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B*, pages 259–302, 1986.

[17] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 127–134, 2003.

[18] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[19] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings in Computational Statistics*, pages 177–186. 2010.

[20] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, England, 2004.

[21] Christopher JC Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11:23–581, 2010.

[22] Ajay Byanjankar, Markku Heikkilä, and Jozsef Mezei. Predicting credit risk in peer-to-peer lending: A neural network approach. In *IEEE Symposium Series on Computational Intelligence*, pages 719–725. IEEE, 2015.

[23] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[24] Ayse Can. The measurement of neighborhood dynamics in urban house prices. *Economic Geography*, 66(3):254–272, 1990.

[25] Bradford Case, John Clapp, Robin Dubin, and Mauricio Rodriguez. Modeling spatial and temporal house price patterns: A comparison of four models. *The Journal of Real Estate Finance and Economics*, 29(2):167–191, 2004.

[26] Bradford Case, Henry O Pollakowski, and Susan M Wachter. On choosing among house price index methodologies. *Real Estate Economics*, 19(3):286–307, 1991.

[27] Karl E Case and Robert J Shiller. The efficiency of the market for single-family homes. *The American Economic Review*, pages 125–137, 1989.

[28] Karl E Case, Robert J Shiller, et al. Prices of single-family homes since 1970: new indexes for four cities. *New England Economic Review*, (Sep):45–56, 1987.

[29] Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1-2):4–28, 2008.

[30] Bisheng Chen, Jingdong Wang, Qinghua Huang, and Tao Mei. Personalized video recommendation through tripartite graph propagation. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1133–1136. ACM, 2012.

[31] Chen Cheng, Fen Xia, Tong Zhang, Irwin King, and Michael R Lyu. Gradient boosting factorization machines. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 265–272. ACM, 2014.

[32] Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. A neural network approach to ordinal regression. In *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1279–1284. IEEE, 2008.

[33] Mikhail Chernov, Brett R Dunn, and Francis A Longstaff. Macroeconomic-driven prepayment risk and the valuation of mortgage-backed securities. *The Review of Financial Studies*, 31(3):1132–1183, 2017.

[34] Sumit Chopra, Trivikraman Thampy, John Leahy, Andrew Caplin, and Yann LeCun. Discovering the hidden structure of house prices with a non-parametric latent manifold model. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 173–182. ACM, 2007.

[35] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[36] Kristof Coussement and Dirk Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, 34(1):313–327, 2008.

[37] Theodore M Crone et al. House prices and the quality of public schools: what are we buying? *Business Review*, (Sep):3–14, 1998.

[38] Andrea Dal Pozzolo, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In *IEEE Symposium Series on Computational Intelligence*, pages 159–166. IEEE, 2015.

[39] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240. ACM, 2006.

[40] Karolien De Bruyne and Jan Van Hove. Explaining the spatial variation in housing prices: an economic geography approach. *Applied Economics*, 45(13):1673–1689, 2013.

[41] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, pages 1–38, 1977.

[42] Dingxiong Deng, Cyrus Shahabi, Ugur Demiryurek, Linhong Zhu, Rose Yu, and Yan Liu. Latent space model for road networks to predict time-varying traffic. *arXiv preprint arXiv:1602.04301*, 2016.

[43] Yue Deng, Feng Bao, Youyong Kong, Zhiquan Ren, and Qionghai Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664, 2017.

[44] Zhengyu Deng, Jitao Sang, and Changsheng Xu. Personalized video recommendation based on cross-platform user modeling. In *IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2013.

[45] Patrik D'haeseleer. What are dna sequence motifs? *Nature Biotechnology*, 24(4):423–425, 2006.

[46] Paramita Dhar and Stephen L Ross. School district quality and property values: Examining differences along school district boundaries. *Journal of Urban Economics*, 71(1):18–25, 2012.

[47] Gang Dong, Kin Keung Lai, and Jerome Yen. Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science*, 1(1):2463–2468, 2010.

[48] Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[49] Mohammed Abdul Haque Farquad, Vadlamani Ravi, and S Bapi Raju. Churn prediction using comprehensible support vector machine: An analytical crm application. *Applied Soft Computing*, 19:31–40, 2014.

[50] Mi Fei and Dit-Yan Yeung. Temporal models for predicting student dropout in massive open online courses. In *IEEE International Conference on Data Mining Workshop*, pages 256–263. IEEE, 2015.

[51] Suhua Feng, Shawn J Cokus, Xiaoyu Zhang, Pao-Yang Chen, Magnolia Bostick, Mary G Goll, Jonathan Hetzel, Jayati Jain, Steven H Strauss, Marnie E Halpern, et al. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, 107(19):8689–8694, 2010.

[52] Francisco Fernández-Navarro, Pedro Antonio Gutiérrez, Cásar Hervás-Martánez, and Xin Yao. Negative correlation ensemble learning for ordinal regression. *IEEE Transactions on Neural Networks and Learning Systems*, 24(11):1836–1849, 2013.

[53] Francisco Fernández-Navarro, Annalisa Riccardi, and Sante Carloni. Ordinal neural networks without iterative tuning. *IEEE Transactions on Neural Networks and Learning Systems*, 25(11):2075–2085, 2014.

[54] Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

[55] Yanjie Fu, Yong Ge, Yu Zheng, Zijun Yao, Yanchi Liu, Hui Xiong, and Jing Yuan. Sparse real estate ranking with online user reviews and offline moving behaviors. In *IEEE International Conference on Data Mining*, pages 120–129. IEEE, 2014.

[56] Yanjie Fu, Guannan Liu, Spiros Papadimitriou, Hui Xiong, Yong Ge, Hengshu Zhu, and Chen Zhu. Real estate ranking via mixed land-use latent models. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 299–308. ACM, 2015.

[57] Yanjie Fu, Hui Xiong, Yong Ge, Zijun Yao, Yu Zheng, and Zhi-Hua Zhou. Exploiting geographic dependencies for real estate appraisal: a mutual perspective of ranking and clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1047–1056. ACM, 2014.

[58] Yanjie Fu, Hui Xiong, Yong Ge, Yu Zheng, Zijun Yao, and Zhi-Hua Zhou. Modeling of geographic dependencies for real estate ranking. *ACM Transactions on Knowledge Discovery from Data*, 11(1):11, 2016.

[59] Ye Fu, Guan-Zheng Luo, Kai Chen, Xin Deng, Miao Yu, Dali Han, Ziyang Hao, Jianzhao Liu, Xingyu Lu, Louis C Doré, et al. N 6-methyldeoxyadenosine marks active transcription start sites in chlamydomonas. *Cell*, 161(4):879–892, 2015.

[60] Ye Fu, Guan-Zheng Luo, Kai Chen, Xin Deng, Miao Yu, Dali Han, Ziyang Hao, Jianzhao Liu, Xingyu Lu, Louis C Doré, et al. N6-methyldeoxyadenosine marks active transcription start sites in chlamydomonas. *Cell*, 161(4):879–892, 2015.

[61] Sheng Gao, Dai Zhang, Honggang Zhang, Chao Huang, Yongsheng Zhang, Jianxin Liao, and Jun Guo. Veclp: A realtime video recommendation system for live

tv programs. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[62] Alan E Gelfand, Mark D Ecker, John R Knight, and C. F. Sirmans. The dynamics of location in home price. *The Journal of Real Estate Finance and Economics*, 29(2):149–166, 2004.

[63] Karen Gibler and Susan Nelson. Consumer behavior applications to real estate education. *Journal of Real Estate Practice and Education*, 6(1):63–83, 2003.

[64] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[65] William N Goetzmann and Liang Peng. The bias of the rsr estimator and the accuracy of some alternatives. *Real Estate Economics*, 30(1):13–39, 2002.

[66] Allen C Goodman. Hedonic prices, price indices and housing markets. *Journal of Urban Economics*, 5(4):471–484, 1978.

[67] Peter J Green and Sylvia Richardson. Hidden markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1055–1070, 2002.

[68] Eric Lieberman Greer, Mario Andres Blanco, Lei Gu, Erdem Sendinc, Jianzhao Liu, David Aristizábal-Corrales, Chih-Hung Hsu, L Aravind, Chuan He, and Yang Shi. Dna methylation on n 6-adenine in c. elegans. *Cell*, 161(4):868–878, 2015.

[69] Zuguang Gu, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. circlize implements and enhances circular visualization in r. *Bioinformatics*, 30(19):2811–2812, 2014.

[70] Sherif Halawa, Daniel Greene, and John Mitchell. Dropout prediction in moocs using learner activity features. *Experiences and Best Practices in and around MOOCs*, 7:3–12, 2014.

[71] James W Hardin and Joseph M Hilbe. *Generalized estimating equations*. Wiley Online Library, 2003.

[72] Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996.

[73] Donald R Haurin and David Brasington. School quality and real house prices: Inter- and intrametropolitan effects. *Journal of Housing Economics*, 5(4):351–368, 1996.

[74] Kathy J Hayes and Lori L Taylor. Neighborhood school characteristics: what signals quality to homebuyers? *Economic Review-Federal Reserve Bank of Dallas*, pages 2–9, 1996.

[75] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell*, 38(4):576–589, 2010.

[76] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. In *The Ninth International Conference on Artificial Neural Networks*, pages 97–102. IET, 1999.

[77] Michal Herzenstein, Rick L Andrews, Uptal Dholakia, and Evgeny Lyandres. The democratization of personal consumer loans? determinants of success in online peer-to-peer lending communities. *Boston University School of Management Research Paper*, 14(6), 2008.

[78] Holger Heyn and Manel Esteller. An adenine code for dna: a second life for n6-methyladenine. *Cell*, 161(4):710–713, 2015.

[79] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[80] Jennifer Jellison Holme. Buying homes, buying schools: School choice and the social construction of school quality. *Harvard Educational Review*, 72(2):177–206, 2002.

[81] Bingquan Huang, Mohand Tahar Kechadi, and Brian Buckley. Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1):1414–1425, 2012.

[82] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

[83] Adnan Idris, Asifullah Khan, and Yeon Soo Lee. Genetic programming and adaboosting based churn prediction for telecom. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 1328–1332. IEEE, 2012.

[84] Guoli Ji, Xiaohui Wu, Yingjia Shen, Jiangyin Huang, and Qingshun Quinn Li. A classification-based prediction model of messenger rna polyadenylation sites. *Journal of Theoretical Biology*, 265(3):287–296, 2010.

[85] Jiming Jiang. *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media, 2007.

[86] Shan Jiang, Joseph Ferreira, and Marta C González. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, pages 1–33, 2012.

[87] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[88] Thomas J Kane, Stephanie K Riegg, and Douglas O Staiger. School quality, neighborhoods, and housing prices. *American Law and Economics Review*, 8(2):183–212, 2006.

[89] Zolidah Kasiran, Zaidah Ibrahim, and Muhammad Syahir Mohd Ribuan. Mobile phone customers churn prediction using elman and jordan recurrent neural network. In *The 7th International Conference on Computing and Convergence Technology*, pages 673–678. IEEE, 2012.

[90] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *The 3rd International Conference for Learning Representations*, 2014.

[91] Yehuda Koren. The bellkor solution to the netflix grand prize. *Netflix Prize Documentation*, 81, 2009.

[92] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[93] Dongwon Lee, David U Gorkin, Maggie Baker, Benjamin J Strober, Alessandro L Asoni, Andrew S McCallion, and Michael A Beer. A method to predict the impact of regulatory variants from dna sequence. *Nature Genetics*, 47(8):955, 2015.

[94] Elisa T Lee and John Wang. *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons, 2003.

[95] Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems*, pages 865–872, 2007.

[96] Stan Z Li. *Markov random field modeling in image analysis.* Springer Science & Business Media, 2009.

[97] Defu Lian, Cong Zhao, Xing Xie, Guangzhong Sun, Enhong Chen, and Yong Rui. Geomf: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 831–840, 2014.

[98] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[99] Bang Liu, Borislav Mavrin, Di Niu, and Linglong Kong. House price modeling over heterogeneous regions with hierarchical spatial functional analysis. In *Proceedings of the 16th International Conference on Data Mining*, pages 1047–1052. IEEE, 2016.

[100] Bin Liu, Yanjie Fu, Zijun Yao, and Hui Xiong. Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1043–1051. ACM, 2013.

[101] De Liu, Daniel Brass, Yong Lu, and Dongyu Chen. Friendships in online peer-to-peer lending: Pipes, prisms, and relational herding. *Management Information Systems Quarterly*, 39(3):729–742, 2015.

[102] Jianzhao Liu, Yuanxiang Zhu, Guan-Zheng Luo, Xinxia Wang, Yanan Yue, Xiaona Wang, Xin Zong, Kai Chen, Hang Yin, Ye Fu, et al. Abundant dna 6ma methylation during early embryogenesis of zebrafish and pig. *Nature Communications*, 7, 2016.

[103] Junxiang Lu. Predicting customer churn in the telecommunications industry—-an application of survival analysis modeling using sas. *Proceedings of the 26th Annual SAS Users Group International Conference*, pages 114–27, 2002.

[104] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications*, 390(6):1150–1170, 2011.

[105] Chunyu Luo, Hui Xiong, Wenjun Zhou, Yanhong Guo, and Guishi Deng. Enhancing investment decisions in p2p lending: an investor composition perspective. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 292–300. ACM, 2011.

[106] Guan-Zheng Luo, Mario Andres Blanco, Eric Lieberman Greer, Chuan He, and Yang Shi. Dna n6-methyladenine: a new epigenetic mark in eukaryotes? *Nature Reviews Molecular Cell Biology*, 16(12):705, 2015.

[107] Yuanhua Lv, Taesup Moon, Pranam Kolari, Zhaohui Zheng, Xuanhui Wang, and Yi Chang. Learning to model relatedness for news recommendation. In *Proceedings of the 20th International Conference on World Wide Web*, pages 57–66. ACM, 2011.

[108] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning*, volume 30, 2013.

[109] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[110] Alireza Mahani and Mansour Sharabiani. *Bayesian, and non-Byesian, cause-specific competing-risk analysis for parametric and non-parametric survival functions: The R Package CFC*. PhD thesis, 2015.

[111] Mercy Marimo. *Survival analysis of bank loans and credit risk prognosis*. PhD thesis, 2015.

[112] Yasuko Matsubara, Yasushi Sakurai, Willem G Van Panhuis, and Christos Faloutsos. Funnel: automatic mining of spatially coevolving epidemics. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 105–114. ACM, 2014.

[113] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta-Protein Structure*, 405(2):442–451, 1975.

[114] Clark L Maxam and Michael LaCour-Little. Applied nonparametric regression techniques: estimating prepayments on fixed-rate mortgage-backed securities. *The Journal of Real Estate Finance and Economics*, 23(2):139–160, 2001.

[115] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 121–128, 2008.

[116] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

[117] Charles E McCulloch and John M Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2001.

[118] Richard Meese and Nancy Wallace. Nonparametric estimation of dynamic hedonic price models and the construction of residential housing price indices. *Real Estate Economics*, 19(3):308–332, 1991.

[119] Michael C Mozer, Richard Wolniewicz, David B Grimes, Eric Johnson, and Howard Kaushansky. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, 11(3):690–696, 2000.

[120] Chaitra H Nagaraja, Lawrence D Brown, and Linda H Zhao. An autoregressive approach to house price modeling. *The Annals of Applied Statistics*, pages 124–149, 2011.

[121] Saurabh Nagrecha, John Z Dillon, and Nitesh V Chawla. Mooc dropout prediction: lessons learned from making pipelines interpretable. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 351–359. IW3C2, 2017.

[122] Shyam V Nath and Ravi S Behara. Customer churn analysis in the wireless industry: A data mining approach. In *Proceedings-annual Meeting of the Decision Sciences Institute*, pages 505–510, 2003.

[123] John A Nelder and R Jacob Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

[124] Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

[125] Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, and Yong Shi. Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, 38(12):15273–15285, 2011.

[126] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4920–4928, 2016.

[127] Kelley Pace, Ronald Barry, John M Clapp, and Mauricio Rodriquez. Spatiotemporal autoregressive models of neighborhood effects. *The Journal of Real Estate Finance and Economics*, 17(1):15–33, 1998.

[128] Kelley Pace, Ronald Barry, Otis W Gilley, and CF Sirmans. A method for spatial–temporal forecasting with an application to real estate prices. *International Journal of Forecasting*, 16(2):229–246, 2000.

[129] Seung-Taek Park and Wei Chu. Pairwise preference regression for cold-start recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*, pages 21–28. ACM, 2009.

[130] Steven Peterson and Albert Flanagan. Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, 31(2):147–164, 2009.

[131] Michael JD Powell. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 2009.

[132] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-fifth Conference on Uncertainty in Artificial Intelligence*, pages 452–461. AUAI Press, 2009.

[133] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[134] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[135] Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc. Buffalo NY, 1961.

[136] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California University San Diego La Jolla Institute for Cognitive Science, 1985.

[137] Laura M Sangalli, James O Ramsay, and Timothy O Ramsay. Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):681–703, 2013.

[138] Eduardo S Schwartz and Walter N Torous. Mortgage prepayment and default decisions: A poisson regression approach. *Real Estate Economics*, 21(4):431–449, 1993.

[139] Chun-Wei Seah, Ivor W Tsang, and Yew-Soon Ong. Transductive ordinal regression. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1074–1086, 2012.

[140] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In *Recommender Systems Handbook*, pages 257–297. Springer, 2011.

[141] Anuj Sharma, Dr Panigrahi, and Prabin Kumar. A neural network based approach for predicting customer churn in cellular network services. *International Journal of Computer Applications*, 27(11):26–31, 2011.

[142] Amnon Shashua and Anat Levin. Ranking with large margin principle: Two approaches. *Advances in Neural Information Processing Systems*, pages 961–968, 2003.

[143] Shashi Shekhar, Zhe Jiang, Reem Ali, Emre Eftelioglu, Xun Tang, Venkata Gunturi, and Xun Zhou. Spatiotemporal data mining: a computational perspective. *ISPRS International Journal of Geo-Information*, 4(4):2306–2338, 2015.

[144] Robert J Shiller. Arithmetic repeat sales price estimators. *Journal of Housing Economics*, 1(1):110–126, 1991.

[145] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[146] Justin Sirignano, Apaar Sadhwani, and Kay Giesecke. Deep learning for mortgage risk. *arXiv preprint arXiv:1607.02470*, 2016.

[147] Tony E Smith and Peggy Wu. A spatio-temporal model of housing prices based on individual sales transactions over time. *Journal of Geographical Systems*, 11(4):333, 2009.

[148] Guojie Song, Dongqing Yang, Ling Wu, Tengjiao Wang, and Shiwei Tang. A mixed process neural network and its application to churn prediction in mobile communications. In *6th IEEE International Conference on Data Mining Workshops*, pages 798–802. IEEE, 2006.

[149] Philip Spanoudes and Thomson Nguyen. Deep learning in customer churn prediction: Unsupervised feature learning on abstract company independent feature vectors. *arXiv preprint arXiv:1703.03869*, 2017.

[150] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[151] Fei Tan, Chaoran Cheng, and Zhi Wei. Modeling real estate for school district identification. In *16th International Conference on Data Mining*, pages 1227–1232. IEEE, 2016.

[152] Fei Tan, Chaoran Cheng, and Zhi Wei. Time-aware latent hierarchical model for predicting house prices. In *16th International Conference on Data Mining*, pages 1111–1116. IEEE, 2017.

[153] Fei Tan, Chaoran Cheng, and Zhi Wei. Modeling and elucidation of housing price. *Data Mining and Knowledge Discovery*, pages 1–27, 2019.

[154] Fei Tan, Kuang Du, Zhi Wei, Haoran Liu, Chenguang Qin, and Ran Zhu. Modeling item-specific effects for video click. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 639–647. SIAM, 2018.

[155] Fei Tan, Xiurui Hou, Jie Zhang, Zhi Wei, and Zhenyu Yan. A deep learning approach to competing risks representation in peer-to-peer lending. *IEEE Transactions on Neural Networks and Learning Systems.*

[156] Fei Tan, Xiurui Hou, Jie Zhang, Zhi Wei, Zhenyu Yan, and Shih-Chuan Weng. A novel risk assessment scheme and practice for peer-to-peer lending. In *ACM SIGKDD Workshop of Data Science in Fintech*. ACM, 2018.

[157] Fei Tan, Yongxiang Xia, and Boyao Zhu. Link prediction in complex networks: a mutual information perspective. *PLoS ONE*, 9(9):e107056, 2014.

[158] Thaddeus Tarpey and Eva Petkova. Latent regression analysis. *Statistical Modelling*, 10(2):133–158, 2010.

[159] Laura O Taylor. The hedonic method. In *A Primer on Nonmarket Valuation*, pages 331–393. 2003.

[160] Chih-Fong Tsai and Yu-Hsin Lu. Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, 36(10):12547–12553, 2009.

[161] Thanasis Vafeiadis, Konstantinos I Diamantaras, George Sarigiannidis, and K Ch Chatzisavvas. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55:1–9, 2015.

[162] Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211–229, 2012.

[163] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

[164] Aurore Voldoire, E Sanchez-Gomez, D Salas y Mélia, B Decharme, Christophe Cassou, S Sénési, Sophie Valcke, I Beau, A Alias, M Chevallier, et al. The cnrm-cm5. 1 global climate model: description and basic evaluation. *Climate Dynamics*, 40(9-10):2091–2121, 2013.

[165] Slobodan Vucetic and Zoran Obradovic. A regression-based approach for scaling-up personalized recommender systems in e-commerce. *WEBKDD00*, 2000.

[166] Artit Wangperawong, Cyrille Brun, Olav Laudy, and Rujikorn Pavasuthipaisit. Churn analysis using deep convolutional neural networks and autoencoders. *arXiv preprint arXiv:1604.05377*, 2016.

[167] Chih-Ping Wei and I-Tang Chiu. Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications*, 23(2):103–112, 2002.

[168] Zhi Wei and Hongzhe Li. A markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23(12):1537–1544, 2007.

[169] Robert B Woodruff. Customer value: the next source for competitive advantage. *Journal of the Academy of Marketing Science*, 25(2):139–153, 1997.

[170] Tao P Wu, Tao Wang, Matthew G Seetin, Yongquan Lai, Shijia Zhu, Kaixuan Lin, Yifei Liu, Stephanie D Byrum, Samuel G Mackintosh, Mei Zhong, et al. Dna methylation on n6-adenine in mammalian embryonic stem cells. *Nature*, 532(7599):329–333, 2016.

[171] Yaya Xie, Xiu Li, EWT Ngai, and Weiyun Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3):5445–5449, 2009.

[172] Zijun Yao, Yanjie Fu, Bin Liu, and Hui Xiong. The impact of community safety on house ranking. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 459–467. SIAM, 2016.

[173] Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *SIGKDD on Knowledge Discovery and Data Mining*, pages 186–194. ACM, 2012.

[174] Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27(3):712–725, 2015.

[175] Scott L Zeger, Kung-Yee Liang, and Paul S Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.

[176] Haoyang Zeng and David K Gifford. Predicting the impact of non-coding variants on dna methylation. *Nucleic Acids Research*, 2017.

[177] Aijun Zhang. *Statistical Methods in Credit Risk Modeling*. PhD thesis, The University of Michigan, 2009.

[178] Guoqiang Zhang, Hua Huang, Di Liu, Ying Cheng, Xiaoling Liu, Wenxin Zhang, Ruichuan Yin, Dapeng Zhang, Peng Zhang, Jianzhao Liu, et al. N 6-methyladenine dna modification in drosophila. *Cell*, 161(4):893–906, 2015.

[179] Guoqiang Zhang, Hua Huang, Di Liu, Ying Cheng, Xiaoling Liu, Wenxin Zhang, Ruichuan Yin, Dapeng Zhang, Peng Zhang, Jianzhao Liu, et al. N6-methyladenine dna modification in drosophila. *Cell*, 161(4):893–906, 2015.

[180] Harry Zhang. The optimality of naive bayes. *American Association for Artificial Intelligence*, 1(2):3, 2004.

[181] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

[182] Hongke Zhao, Qi Liu, Guifeng Wang, Yong Ge, and Enhong Chen. Portfolio selections in p2p lending: A multi-objective perspective. In *Proceedings of the 22nd SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2075–2084. ACM, 2016.

[183] Hongke Zhao, Le Wu, Qi Liu, Yong Ge, and Enhong Chen. Investment recommendation in p2p lending: A portfolio perspective with risk management. In *2014 IEEE International Conference on Data Mining*, pages 1109–1114. IEEE, 2014.

[184] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology*, 5(3):38, 2014.

[185] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931, 2015.

[186] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In *Proceedings of the 20th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 135–144. ACM, 2014.

[187] Zhi-Hua Zhou and Ji Feng. Deep forest: Towards an alternative to deep neural networks. *arXiv preprint arXiv:1702.08835*, 2017.

[188] Hengshu Zhu, Hui Xiong, Fangshuang Tang, Qi Liu, Yong Ge, Enhong Chen, and Yanjie Fu. Days on market: Measuring liquidity in real estate markets. In *Proceedings of the 22nd SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 393–402. ACM, 2016.