Days on Market: Measuring Liquidity in Real Estate Markets

Hengshu Zhu[1], Hui Xiong[2], Fangshuang Tang[3], Qi Liu[3], Yong Ge[4], Enhong Chen[3], Yanjie Fu[5]

[1]Baidu Research-Big Data Lab, [2]Rutgers University, [3]University of Science and Technology of China,
[4]University of Arizona, [5]Missouri University of Science and Technology

[1]zhuhengshu@baidu.com, [2]hxiong@rutgers.edu, [3]{fstang,qiliuql,cheneh}@ustc.edu.cn,
[4]ygestrive@gmail.com, [5]yanjiefoo@gmail.com

ABSTRACT

Days on Market (DOM) refers to the number of days a property is on the active market, which is an important measurement of market liquidity in real estate industry. Indeed, at the micro level, DOM is not only a special concern of house sellers, but also a useful indicator for potential buyers to evaluate the popularity of a house. At the macro level, DOM is an important indicator of real estate market status. However, it is very challenging to measure DOM, since there are a variety of factors which can impact on the DOM of a property. To this end, in this paper, we aim to measure real estate liquidity by examining multiple factors in a holistic manner. A special goal is to predict the DOM of a given property listing. Specifically, we first extract key features from multiple types of heterogeneous real estate-related data, such as house profiles and geo-social information of residential communities. Then, based on these features, we develop a multi-task learning based regression approach for predicting the DOM of real estates. This approach can effectively learn district-aware models for different property listings by considering multiple factors. Finally, we conduct extensive experiments on real-world real estate data collected in Beijing and develop a prototype system for practical use. The experimental results clearly validate the effectiveness of the proposed approach for measuring liquidity in real estate markets.

Keywords

Days on Market, Real Estate, Multi-Task Learning

1. INTRODUCTION

Real estate is an important investment option in many countries and has traditionally outperformed the stock market [1]. However, real estate has limited liquidity compared to other investments [28]. Days on Market (DOM) is an important measurement for market liquidity in real estate industry, as it refers to the number of days a property is on the active market. At the micro level, DOM is a special concern of property sellers because real estate is highly cash flow dependent [1]. Also, it is a useful indicator for potential buyers to evaluate the popularity of the estate [25, 28]. Moreover, at the macro level, DOM is an important indicator of the liquidity of the real estate market and shows the level of risk associated with real estate investments.

In the literature, there are a number of studies about DOM related topics [27, 18, 25, 28, 16]. However, most of existing works focus on developing interpretable models for discerning the relationship between DOM and marketing features, such as listing/selling prices. Less efforts have been made for developing predictive methods for measuring DOM, which is very critical for many parties in real estate industry. To this end, in this paper, we aim to measure real estate liquidity by examining multiple underlying factors in a holistic manner. Along this line, an important objective is to model and predict the DOM of a given property listing. The right prediction of DOM could provide valuable insights for both sellers and buyers of real estates and enable the information transparency between sellers and buyers. For example, a seller would know how long it will probably take to sell the house at a certain price using our system. Therefore, the seller could adjust the price to influence the expected DOM. If it turns out that well-decorated houses are more popular (with short DOM) on the market, the seller could also decorate the house to make it more popular. In addition, the developed system can help identify the deceptive manipulations [28] on DOM, and thus could lead to the healthy development of the real estate industry.

However, it is very challenging to predict the DOM of a property because DOM is potentially affected by many factors, such as price, location, and the year of completion in a complicated and involved way. To this end, we first collect and investigate a variety of real estate-related data including transaction records, estate profiles and geo-social information of residential communities. By carefully studying our real-world data, we find that some of these factors are only weakly correlated with DOM. For instance, although some attributes from property profiles (e.g., unit price) can be regarded as intuitively discriminant information for DOM prediction, it turns out that they have limited prediction power for DOM because most property listings within the same urban district have very similar profile attributes. Therefore, we extract various contextual features for establishing the key factors that affect the sale of real estates. These feature include the heterogeneous house profile and the geo-

KDD '16, August 13-17, 2016, San Francisco, CA, USA

© 2016 ACM. ISBN 978-1-4503-4232-2/16/08…$15.00

DOI: http://dx.doi.org/10.1145/2939672.2939686

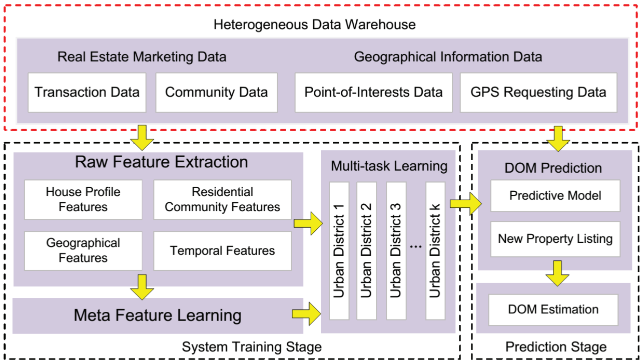


Figure 1: An overview of the framework.

social information of residential communities, such as the number of available facilities (e.g., school and hospital) in the neighborhood, population density, news, and the average monthly DOM of housing sales in the same community. Then, to make full use of all relevant features, we develop a multi-task learning based regression approach for DOM prediction, which can effectively learn district-aware prediction models capturing the geographical correlations of real estates. Figure 1 demonstrates the work flow overview of our approach. Finally, we evaluate our approach by conducting extensive experiments with a large amount of real-world estate related data in Beijing, which include 10K+ residential communities, 145K+ transaction records, 740K+ real estate news, 500K+ points of interests (POIs), and 59M+ GPS requests. Experimental results have demonstrated the effectiveness of the proposed method. In addition, we develop a prototype system for practical use, which clearly shows the value of the proposed decision support system.

Overview. The remainder of this paper is organized as follows. In Section 2, we introduce the details of large-scale real estate data sets. Section 3 presents how to extract contextual features for DOM prediction. In Section 4, we introduce the technical details of our DOM prediction model by exploiting multi-task learning. In Section 5, we report the experimental results of DOM prediction. Section 6 provides a brief review of related works. Finally, in Section 7, we conclude the paper.

2. DATA DESCRIPTION

In this section, we introduce a variety of real-estate related data sets that we have used in developing our multi-task learning based regression method and the DOM decision support system. In summary, Table 2 shows the statistics of our real-world data sets.

2.1 Real Estate Marketing Data

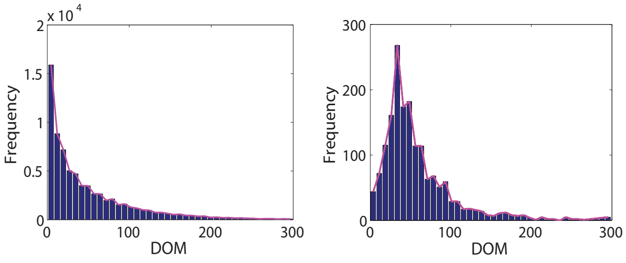
In this study, we use two sets of marketing data collected from a major commercial real estate agency in China.

The first one is a long-term real estate transaction data set, which contains 145,932 transaction records from October 2011 to November 2013 in Beijing. Figure 2 (a) shows the distribution of the number of transactions with respect to different length of DOM. We can observe that most of house listings only have very short DOM, which clearly indicates the prosperity of real estate markets in China. Indeed, this distribution is quite different from that of another data set from Realtor [3] in US, shown in Figure 2 (b). Therefore, we believe, this study can provide valuable insights into Chinese real estate markets.

Table 2: A Summary of Data Statistics.

Data Type	Amount
Raw Transactions	145,932
Community Profiles	10,425
Real Estate News	740,434
Point-of-Interests	510,747
GPS Requesting	59,638,947

Each transaction record in our data set also contains the profile of the listed house, including the residential community the house belongs to [1], DOM, price, location, room number, area, list date, whether it is free of sales tax, orientation, the realtor who facilitated this transaction, etc. Specifically, the top part of Table 1 shows some basic statistics of the above attributes in our data set. Figure 3 shows the distributions of DOM, unit price and transaction number with respect to list time in our data set. From these figures, we can have some very interesting findings. First, the average unit price has grown steadily in Beijing. Second, there is a burst of transaction volume in March 2013, which is due to the "*five policies and measures to regulate real estate market*" by Chinese government [4]. Furthermore, Figure 4 shows the heat map of the geographical distributions of DOM, unit price and transaction numbers in our data set. We can observe that the DOM distribution is relatively even, and the locations with high unit prices usually have a low number of transaction records, which is most prominent when comparing the upper left part of Figure 4 (b), (c).



(a) Our Data Set (b) Realtor Data Set

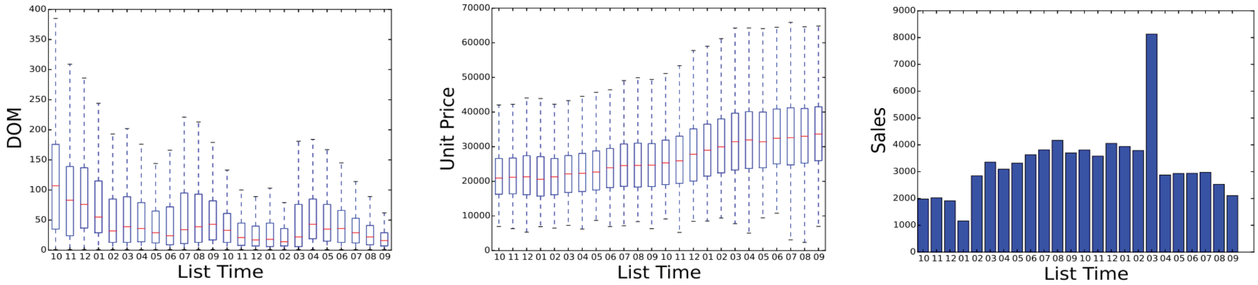
Figure 2: The DOM distribution.

Another market data set includes the profile information of 10,425 residential communities in Beijing, including building number, greening rate, plot ratio, completion year, etc. Particularly, since the real estate market is usually influenced by public opinions, we also collected a large number of real estate news from a variety of portal websites. By using a commercial API of named entity recognition [2], we finally obtained 740,434 news that can be linked to the communities in our data set. Detailed data descriptions are shown at the middle part of Table 1. Moreover, the data set also contains the neighboring facility information of each community, including Transport (e.g., bus stations), School, Hospital, Entertainment (e.g., cinemas), Shopping, Scenery and Unpleasant Facilities (e.g., factories). Some statistics are shown at the bottom part of Table 1, and Figure 3 (d) shows the heat map of the geographical distribution of residential communities.

[1] In the cities of China, a house usually belongs to a specific residential community.

Table 1: The statistics of some basic attributes of our real estate marketing data sets.

Data Type	Attributes	Min	Max	Mean	Median	Description
Transaction Profile	Room Number	1	8	2	2	Number of rooms in the house
	Price (10^4 RMB)	10	4,380	224	196	Total price of the house
	Unit Price (RMB)	1,422	380,000	28,288	26,316	Price of the house per square meter
	Area (m^2)	5	996	82	74	Area of the house
	DOM	1	737	53	30	Days on Market of the house
Community Profile	Building Num	1	363	22	17	Number of buildings in the community
	Greening Rate	0.00	0.89	0.31	0.31	Greening rate of the community
	Plot Ratio	0.1	82.0	2.4	2.1	Plot ratio of the community
	News Num	0	2,396	71	32	Number of news published about the community
	Completion Year	1963	2015	2002	2003	Completion year of the community
Neighboring Facilities	Transport	0	13	4	3	Number of transport stations nearby
	School	0	61	15	12	Number of schools nearby
	Hospital	0	53	15	13	Number of hospitals nearby
	Entertainment	0	71	9	5	Number of entertainment facilities nearby
	Shopping	0	45	11	11	Number of shopping malls nearby
	Scenery	0	10	2	2	Number of sightseeing spots nearby
	Unpleasant	0	20	5	4	Number of unpleasant facilities nearby



(a) DOM

(b) Unit Price

(c) Transaction Number

Figure 3: The distributions of DOM, unit price and transaction number w.r.t list time (grouped by month), ranging from Oct. 2011 to Sept. 2013. For brevity, we only show the "month" in x-axis.

2.2 Geographical Information Data

Other than the transaction data, we also collected two sets of large-scale geographical information from a major commercial online map service provider in China. These data sets contain two kinds of geographical information, namely Points Of Interest (POIs) and GPS requests of mobile users.

Specifically, there are totally 510,747 POIs of Beijing in our data set, which can help us further analyze the functionality and convenience of real estates. For example, more POIs generally indicate more convenient living environment, thus the density of POIs is an important factor in influencing people's buying decisions. Figure 4 (e) shows the heat map of the geographical distributions of residential community and POI in our data set.

In another data set, we have the fine-grained daily GPS requesting records of mobile users in Beijing. To be specific, in our experiments, we filtered in total 59,638,947 GPS requests of mobile users from 18:00PM to 23:59PM of some workdays in Beijing. The heat map of the geographical distribution of GPS request is shown in Figure 4 (f). Since most people would stay at home during this time period, we believe that this distribution could roughly reflect the population distribution of Beijing. These results can help us evaluate the occupancy rate and popularity of each residential community.

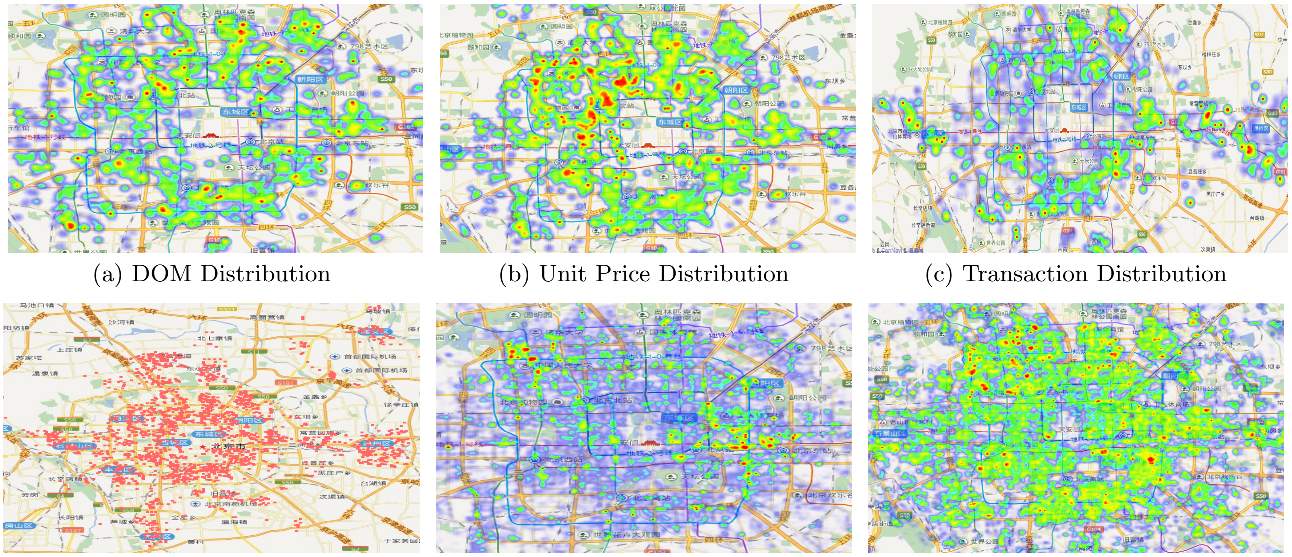
3. FEATURE EXTRACTION

In this section, we introduce the feature extraction from our data for DOM prediction. Specifically, we group all features into five categories, namely *house profile features*, *residential community features*, *geographical features*, *temporal features*, and *meta features*. While some are simple transformations of house/community attributes, others are implicit features that require the mining of raw data.

3.1 House Profile Features

We obtain 11 house profile features such as area, price, and decoration, which describe the basic characteristics of the house. The details of these features are illustrated in Table 3. Particularly, the feature "Free of Tax" is a dummy variable indicating the seller has to pay sales tax or not. In China, if a residential house is resold by a seller within 5 years since the seller bought it, the seller has to pay sales tax. Otherwise, the seller will be free of sale tax. It is a policy of Chinese government to discourage transactions of residential houses for investment purposes. All of house profile features in Table 3 are directly available from the raw data, except the feature "Historical DOM of Same Realtor", which is used to represent the experience of realtors on DOM. The assumption here is that there exist hard-working or highly skilled realtors who are more likely able to sell a house in a short time, while a less-motivated or inexperienced realtor may spend a long time to find a buyer for the same house. For each transaction, this feature is calculated by averaging the DOM of houses sold by the same realtor.

What should be noted here is that, in the prediction model, we implement classic standardization for all numerical features. And we convert each categorical feature into N binary-valued features using one-hot encoding (i.e., dummy feature), where N is the number of possible values that feature


(a) DOM Distribution (b) Unit Price Distribution (c) Transaction Distribution

(d) Community Distribution (e) POI Distribution (f) GPS Request Distribution

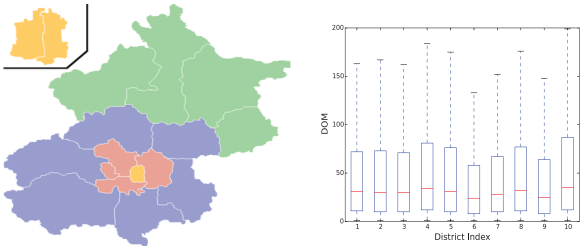
Figure 4: The geographical distributions of DOM, unit price, transaction number, residential community, POIs and GPS requests.

could be. Unless specified otherwise, we employ the same recipe to transform other features in this section.

3.2 Residential Community Features

When buying a property, people usually consider not only the house profile, but also the profile of residential community where the house is located. The features we select to describe the profile of residential communities are illustrated in Table 3. To reflect the unique characteristics of each residential community, for each transaction record, we calculate a feature "Historical DOM of Same Community" to indicate the average DOM of historical house transactions in the same community. Since houses in the same community are usually similar in functionality, it is possible that their DOMs are also similar. In such case, this feature could serve as a useful predictor.

Particularly, "District" is an important characteristic of residential communities. Indeed, each city in China is always segmented into different urban districts for administration, which usually have unique urban functionalities, such as business and education. Therefore, we ague that houses in different districts will have different DOM distributions. For example, in our data, there are transaction records from 10 districts out of the 16 districts in Beijing (e.g., as shown in Figure 5 (a)), and the corresponding DOM distribution with respect to different districts is shown in Figure 5 (b). Besides extracting features from district information for DOM prediction, we also use district information as a criterion to split tasks when performing multi-task learning.



(a) (b)

Figure 5: The DOM distribution with respect to different urban districts.

3.3 Geographical Features

Here we extract several geographical features to capture the spatial characteristics of real estate market. The details are shown in Table 3 and explained as follows.

Surrounding Facilities. The surrounding facilities are important when people consider buying a house, since most of our living services are not directly from the house or the residential community, but from the surrounding facilities. In our data sets, we have seven kinds of facilities, including Transport, School, Hospital, Entertainment, Shopping, Scenery and Unpleasant Facilities. Therefore, for each house, we count the number of each kind of facilities near the house, and use them as geographical features.

Population Density. Indeed, the surrounding population density of house is also useful for DOM prediction, because a residential community that has more people indicates that it is more popular and thus may have smaller DOM. By using the large-scale data set of GPS requests, we can effectively estimate the population density of different areas. Specifically, we first use the widely-adopted grid based method to partition the city area of Beijing into fine-grained grids. Then, we consider the number of requests within each grid as an estimate of population density. For each listed house, we use the density of the grid where it is located as the corresponding geographical feature.

POI Density. As mentioned in Section 2.2, the POI density is an important criterion when people decide to buy the house or not. Therefore, here we extract this geographical feature in a similar way to the population density.

3.4 Temporal Features

To fully explore the temporal characteristics of real estate

Table 3: The description of features.

Feature Type	Feature	Description
House Profile	Room Number	Number of rooms in this house
	Total Price	Total price of the house (10^4 RMB)
	Area	Number of square meters of the house
	Unit Price	Price of the house per square meter (RMB)
	Free of Tax	Dummy variable indicating the house is free of sales tax
	Floor Number	Number of floors of the building where the house is located
	Floor Type	Type of the floor of the house: high, medium, low, or in the basement
	Orientation	Orientation of the house: south, north, east, west, etc.
	Decoration	Type of decoration: well-decorated, simply-decorated, not decorated
	Building Type	Building type: slab-type building, tower building, mixed-type building, etc.
	Historical DOM of Same Realtor	Average DOM of houses employing the same realtor
Residential Community Profile	District	The district of the house: Haidian, Chaoyang, Xicheng, etc
	Completion Year	Completion year of the community
	Greening Rate	Greening rate of the community
	Plot Ratio	Plot ratio of the community
	News Number	Current number of news published about the community
	Historical DOM of Same Community	Average DOM of sold houses in the same community
Geographical Feature	Within School District	Dummy variable indicating whether the house is near school
	Transport	Number of transport stations nearby
	School	Number of schools nearby
	Hospital	Number of hospitals nearby
	Entertainment	Number of entertainment facilities nearby
	Shopping	Number of shopping malls nearby
	Scenery	Number of sightseeing spots nearby
	Unpleasant	Number of unpleasant facilities nearby
	Population Density	Estimated population density within nearest 1 km×1km grid
	POI Density	POI density within nearest 1 km×1km grid
Temporal Feature	List Date	The date of corresponding property (only day and month information is used)
	DOM of Recently Sold Houses	The DOM of the most recent sold house
	Average DOM of Recently Sold Houses	In recent N days, the average DOM of sold houses
	Percentage of Recently Sold Houses	In recent N days, the percent of listed houses that are sold
	DRPAP	Unit price minus the average unit price of sold houses in recent N days
Meta Feature	RF Feature	Results of random trees trained on original data

market, we also extract the following temporal features for DOM prediction. The details of these features are shown in Table 3 and explained as follows.

DOM of Recently Sold Houses. Two houses that are temporally near to each other may have similar DOM values, since they share similar market conditions. To capture such correlation, for each listed house, we retrieve the most recent N (e.g., 5 in our experiments) sold houses in the same community before its list date, and use the DOM's of those N houses as N numeric features. Note that, if we cannot find such N houses in the same community, we directly use the average DOM of recently sold houses as the default value.

Average DOM of Recently Sold Houses. The intuition behind this feature is similar to that of "DOM of Recent Sold", but differs in that it uses fixed time spans. Specifically, for each listed house, we calculate the average DOM of the houses sold in the same community within the past N days. If no house is sold within that period, the average DOM of all sold houses in the same community is used as the default value. Indeed, this feature can be regarded as complementary to the previous feature. In our experiments, we choose N to be 3, 10, 20, 30, 50 to get 5 different features.

Percentage of Recently Sold Houses. To reflect the temperature of real estate market, we also propose to calculate the percentage of recent sold houses. Specifically, for each listed house, we calculate how much percent of the houses listed within past N days has been sold. If there is no houses listed within this period, this feature is set to a default value (i.e., 0.5 in our experiments). In our experiments, we choose N to be 10, 30, 50 to get 3 different features.

Difference of the Recent Price and the Average Price (DRPAP). Although the price of a house has been widely recognized as a critical variable in influencing DOM [8], it often changes over time. Therefore, we believe that the relative price difference is a better predictor of DOM than the absolute price value. Specifically, for each listed house, we calculate the difference between unit price of current transaction record and the average unit price of the houses sold within past N days. If there is no house listed within this period, this feature is set to a default value (i.e., 0 in our experiments). In experiments, we set N to be 10.

When extracting temporal features, we should be very careful not to use future information, especially when we use default average to fill missing features, since this may cause test data to be used in the training process.

3.5 Meta Features

To improve the accuracy of DOM prediction, we also design some meta features by borrowing the idea of ensemble learning, i.e., training a model on the training data, using the model to predict the DOM of all the instances, and using the prediction value of each instance as a feature. Indeed, such meta features can be regarded as the effective substitutes of manual-selected feature combinations, and thus can help to avoid the "Curse of Dimensionality" problem in feature engineering. In this study, we choose the random forest model for generating meta features due to two reasons: 1) random forest model explores random feature combinations, which can generate diverse structures and avoid homogeneity; 2) the superior performance of random forest model on both classification and regression tasks has been well proved by previous research [24]. Specifically, we first choose the model parameters (e.g. number of trees, max depth of trees) using validation set that can produce best results. Then, after fitting the random forest model on our training data, we use the trained decision trees in the forest to predict DOM for each listed house, and append the prediction results as meta features to other features.

4. MULTI-TASK LEARNING FOR DOM PREDICTION

In this section, we introduce our DOM prediction model, which is built upon multi-task learning.

4.1 Motivation

An intuitive idea for DOM prediction is directly training a regression model with historical transaction records. However, as a kind of geographical asset, real estate usually has unique dependency on locations. For example, as discussed in Section 3, a city in China is always partitioned into different urban districts for administration, which usually have unique urban functionalities and result in different DOM distributions. Therefore, a more reasonable solution is to learn different prediction models for different districts. Figure 6 visualizes the coefficient vectors of linear regression models learned from transaction records of different districts. We can observe that, the coefficient vectors do have different characteristics from each other, which clearly indicates the location dependency of real estates. Although learning a unique model for each district seems to be an effective way for DOM prediction, such solution fails to capture the commonality of different districts and hence has inferior generalization ability. In fact, according to the ***Tobler's First Law of Geography*** [26], "*everything is related to everything else, but near things are more related than distant things*". Therefore, in this study we propose to leverage multi-task learning for addressing this challenge, which can be seen as a trade-off between the previously mentioned two "extreme" solutions. In our approach, we can learn different district-aware models for DOM prediction, and control the distance-aware similarities of different models by using different regularization parameters.

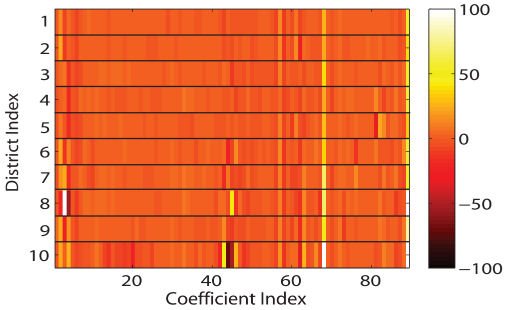


Figure 6: The heat map visualization of coefficient vectors of district-specific linear regression models.

4.2 A DOM Prediction Model

Given a set of N historical transaction records, i.e., $\mathcal{H} \equiv \{(\mathbf{x}_i, y_i)|i = 1, 2, 3, \dots, N\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and y_i represent the extracted feature vector and the DOM of the i-th record, respectively. In our approach, we first divide $\mathcal{H}$ into M disjoint subsets by districts: $\mathcal{H} = \mathcal{H}^1 \cup \mathcal{H}^2 \cdots \cup \mathcal{H}^M$, where $\mathcal{H}^t$ $(1 \leq t \leq M)$ denotes the data set of historical transaction records within the t-th urban district. Then, the problem of DOM prediction can be modeled by minimizing the squared error and regularization term of all the tasks, that is,

$$\min_W f = \sum_{t=1}^{M} ||Y^t - X^t W_{t,:}^T||_2^2 + R(W), \tag{1}$$

where t is the index of task; $Y^t = [y_1^t, \cdot, y_{n_t}^t]^T$; $X^t = [\mathbf{x}_1^t, \cdot, \mathbf{x}_{n_t}^t]^T$; $R(x)$ is the regularization term; W is a coefficient matrix and $W_{t,:}$ is the coefficient vector of regression model for the t-th task. The first term represents the squared error and the second term is used to impose the similarity regularization.

Different choices of regularization terms may reflect different types of task relationships. In this study, we formulate our model by graph regularizer combined with L1 regularizer, as shown in Equation 2.

$$\begin{aligned}\min_W f = &\sum_{t=1}^{M} ||Y^t - X^t W_{t,:}^T||_2^2 + \rho||W||_1 + \\ &\frac{\lambda}{2}\sum_{t_1=1}^{M}\sum_{t_2=1}^{M} S_{t_1 t_2}||W_{t_1,:} - W_{t_2,:}||_2^2, \lambda > 0, \rho > 0,\end{aligned} \tag{2}$$

where S is an $M \times M$ matrix to capture the similarities between different tasks; ρ and λ are hyper-parameters. Specifically, in our model the L1 regularizer controlled by ρ is used to induce a sparse model and avoid over-fitting, and the graph regularizer controlled by λ is used to capture the geographical correlations among tasks, since the tasks are partitioned with respect to urban district. Furthermore, we can find that the graph regularizer is very suitable for representing our idea of trade-off. Specifically, if we omit the L1 regularizer and choose $\lambda = 0$, the formulation is equivalent to training unique prediction model for each district; while if we omit the L1 regularizer and choose $\lambda = +\infty$, the formulation is equal to training a single model on all historical transaction records.

The graph regularizer in Equation 2 can be represented by $\lambda \cdot tr(W^T LW)$, where L is the Laplacian matrix, i.e., $L_{ij} = \delta_{ij} \sum_{k=1}^{M} S_{ik} - S_{ij}$, $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Indeed, the task similarity matrix S can be computed or learned in various ways. In this paper we develop a novel similarity matrix based on the *Tobler's First Law of Geography.* Specifically, we first compute the distance matrix D for all districts, where $D_{i,j}$ is the distance between district i and district j that can be calculated by the average mutual distance of residential communities located in the two districts. If we denote the set of residential communities located in district i and j by C_i and C_j, we have

$$D_{i,j} = \frac{1}{|C_i| \cdot |C_j|} \sum_{c_i \in C_i} \sum_{c_j \in C_j} dist(c_i, c_j), \tag{3}$$

where $c_i \in C_i$ means c_i is a residential community in C_i, and $dist(c_i, c_j)$ is the geographical distance between c_i and c_j. Then, we can compute the similarity matrix S by

$$S = 1 - D/(D_{max} + \delta), \tag{4}$$

where D_{max} is the maximum value in D, and δ is a small positive constant for smoothing.

Note that the first term of f in Equation 2 is a continuous differentiable convex function, and the regularization term is also convex, thus we can use the canonical accelerated gradient method to obtain the global minimum of f. Specifically, let $f = g + h$, where $g = \sum_{t=1}^{M} ||Y^t - X^t W_{t,:}^T||_2^2 + \lambda \cdot tr(W^T LW)$, $h = \rho||W||_1$, we have

$$\begin{aligned}
\nabla g(W)_{ij} = \frac{\partial g}{\partial W_{ij}} &= \frac{\partial ||Y^i - X^i W_{i,:}^T||_2^2}{\partial W_{ij}} + \lambda(LW + L^T W)_{ij} \\
&= \frac{\partial \sum_{k=1}^{K_i} (Y_k^i - X_{k,:}^i W_{i,:}^T)^2}{\partial W_{ij}} + \lambda(LW + L^T W)_{ij} \\
&= \sum_{k=1}^{K} 2(Y_k^i - X_{k,:}^i W_{i,:}^T)(-X_{kj}^i) + \lambda(LW + L^T W)_{ij}.
\end{aligned}$$

The detailed algorithm for learning our prediction model is described in Algorithm 1.

Algorithm 1 Accelerated Gradient Method

```
Input: $S, X, Y, tol, maxIter$
Output: $W$
$W, W_{old} \Leftarrow \mathbf{0}, t \Leftarrow 1.0$
$\beta \in (0, 1)$
for $k = 1$ to $maxIter$ do
    $W_{in} \Leftarrow W + \frac{k-2}{k+1}(W - W_{old})$
    $W_{new} \Leftarrow prox_t(W_{in} - t\nabla g(W_{in}))$
    while $g(W_{new}) > g(W_{in}) + \nabla g(W_{in})^T (W_{new} - W_{in}) + \frac{1}{2t}||W_{new} - W_{in}||_F^2$ do
        $t = \beta t$
        $W_{new} \Leftarrow prox_t(W_{in} - t\nabla g(W_{in}))$
    end while
    $W_{old} \Leftarrow W$
    $W \Leftarrow W_{new}$
    if $\frac{|f(W_{old}) - f(W)|}{|f(W_{old})| + \epsilon} < tol$ then
        Break
    end if
end for
return $W$
```

Note that in Algorithm 1, $prox_t(V) = \arg\min_Z \frac{1}{2t}||V - Z||_2^2 + \rho||Z||_1$ has a closed-form solution, which could be used to design an efficient implementation. The solution is shown as follows.

$$Z_{ij} = \begin{cases} V_{ij} - \rho t, & \text{if } V_{ij} > \rho t; \\ V_{ij} + \rho t, & \text{if } V_{ij} < -\rho t; \\ 0, & \text{otherwise.} \end{cases}$$

5. EXPERIMENTAL RESULTS

In this section, we present experimental results to demonstrate the performances of our DOM prediction approach.

5.1 The Experimental Setup

In our experiments, we removed noisy transaction records of which the DOM or other basic profile information (e.g., price) of property is missing. Meanwhile, we also removed some sparse transaction records if the total number of transaction records in the same residential community is less than a threshold (e.g., 10 in our experiments). To avoid the government-driven (i.e., policy) transactions (i.e., as illustrated in Figure 3 (c)), which may introduce strong bias during model training, we further filtered some transaction records with very short DOM (e.g., 1 day) in March 2013. After the above data pre-processing, this data set has totally 70,149 transaction records remained.

Table 4: Data partitioning.

Data Set	Train	Validation	Test
D#1	9,347 (13.3%)	6,407 (9.1%)	54,395 (77.5%)
D#2	23,192 (33.1%)	7,867 (11.2%)	39,090 (55.7%)
D#3	46,431 (66.2%)	11,915 (17.0%)	11,803 (16.8%)

Due to the temporal property of our transaction data set, we propose to use a pair of dates for splitting the data set into training data, validation data, and test data. Specifically, in our experiment, we choose 3 different pairs of dates for data splitting, i.e., (Apr. 1st, 2012, June 1st, 2012), (Aug. 1st, 2012, Oct. 1st, 2012), and (Feb. 1st, 2013, Apr. 1st, 2013), and thus obtain three evaluation data sets D#1, D#2, and D#3. The statistics of each data set (number of instances and corresponding percentage) are illustrated in Table 4. And all the experiments were conducted on a 2.5GHZ×4-Core CPU, 4G main memory PC with Python 2.7 and Matlab 2012 under Windows 7 64bit system.

In the experiments, our approach is called **Multi-task Linear Regression for DOM prediction (MLR-DOM)**, where all the hyper-parameters are learned by using validation data set.

5.1.1 Evaluation Baselines

To verify the effectiveness of our approach, we also chose several state-of-the-art regression methods as baselines.

- **Linear Regression (LR)**: training a linear regression model for DOM prediction.
- **Ridge**: training a ridge regression model (i.e., LR with L2-norm regularizer) for DOM prediction.
- **Lasso**: training a Lasso regression model (i.e., LR with L1-norm regularizer) for DOM prediction.
- **Location-specific Linear Regression (LsLR)**: training different linear regression models for transaction records in different district.
- **Decision Tree (DT)**: training a CART tree for DOM prediction.
- **Random Forest (RF)**: training a random forest regressor for DOM prediction.
- **Support Vector Regression (SVR)**: training a Support Vector Regression model for DOM prediction.

Note that, all the hyper-parameters of baselines are selected by using validation data set.

5.1.2 Evaluation Metrics

To evaluate the prediction performance of different methods, here we select two widely used metrics from general regression analysis, i.e., the root mean squared error (rMSE) and mean absolute error (MAE), and another widely used metric from multi-task learning, i.e., normalized mean squared error (nMSE) [12] for evaluation.

5.2 Overall Results

Here we present the overall performance comparison between our approach and different baselines. Specifically, Table 5 shows the results of different approaches with respect to different evaluation metrics. From this table, we can have several insightful observations.

Table 5: The performance of different methods.

Data Set	Method	rMSE	nMSE	MAE
D#1	LR	111.2758	4.2616	94.1635
	Ridge	58.4916	1.1775	42.0547
	Lasso	58.6558	1.1841	42.1507
	LsLR	147.2266	7.4599	119.4437
	DT	74.4944	1.9099	56.7929
	RF	76.5393	2.0162	63.0411
	SVR	54.9765	1.0402	43.5489
	MLR-DOM	**54.7480**	**1.0316**	**36.4876**
D#2	LR	66.5383	1.9559	52.1974
	Ridge	48.8114	1.0526	37.6255
	Lasso	49.3662	1.0766	37.8580
	LsLR	115.6633	5.9100	93.3486
	DT	55.5625	1.3639	45.9868
	RF	72.3710	2.3139	61.2322
	SVR	49.4277	1.0793	40.4608
	MLR-DOM	**46.3684**	**0.9498**	**32.7723**
D#3	LR	49.7731	1.5896	38.2208
	Ridge	50.1686	1.6150	39.0059
	Lasso	50.0549	1.6076	38.8514
	LsLR	60.3555	2.3372	41.3725
	DT	39.2197	0.9870	31.8181
	RF	38.7276	0.9624	31.7265
	SVR	39.4896	1.0006	31.6821
	MLR-DOM	**38.5147**	**0.9517**	**30.7996**

First, our approach MLR-DOM consistently outperforms other baselines on all data sets in terms of all evaluation metrics, which clearly validates the effectiveness of our multi-task learning based regression model. Second, the district-aware model LsLR cannot achieve good prediction performance, which is even worse than the basic LR model. It may be because of the imbalanced distribution of training data on different districts. Therefore, using similarity matrix for controlling parameter learning is very important for training district-aware models (i.e., just as our approach MLR-DOM). Third, Ridge and Lasso are two competitive methods on D1 and D2, while DT and RF are two competitive methods on D3. Therefore, although some state-of-the-art methods can perform good on some data sets, their performances are actually not stable or robust.

Furthermore, we have also conducted a series of paired T-test of 0.95 confidence level for the experimental results. The results have shown that the improvements of our approaches to other baselines are all statistically significant.

5.3 Feature Contribution Analysis

To evaluate the effect of different feature integrations, we conducted our multi-task learning approach with three different sets of features, as described below.

- **Raw**: Using raw attributes of house and residential community as features.
- **Raw+ST**: Besides features in **Raw**, geographical and temporal features are also included.
- **Raw+ST+M**: Besides features in **Raw+ST**, meta features are also included.

Table 6: Evaluation of feature combinations.

Data Set	Feature	rMSE	nMSE	MAE
D#1	Raw	55.6912	1.0674	36.5032
	Raw+ST	55.6013	1.0668	36.5616
	Raw+ST+M	**54.7480**	**1.0316**	**36.4876**
D#2	Raw	47.4254	0.9936	34.7932
	Raw+ST	47.4334	0.9881	34.7838
	Raw+ST+M	**46.3684**	**0.9498**	**32.7723**
D#3	Raw	39.2821	0.9900	**30.7882**
	Raw+ST	38.5222	0.9521	30.8136
	Raw+ST+M	**38.5147**	**0.9517**	30.7996

The experimental results are shown in Table 6, where we can have several observation as follows. First, **Raw+ST+M** almost always has the best performance, which indicates the effectiveness of our feature extraction. Second, on D#1 and D#2, adding geographical and temporal features to raw features will not improve the performance much, while adding meta features seems to have produced a noticeable improvement. Third, on D#3 the situation is totally reversed, adding geographical temporal features can improve the prediction performance, while meta features have little effect. Therefore, we can argue that the features are internally correlated, and our model can have the best performance only when combining all kinds of contextual features.

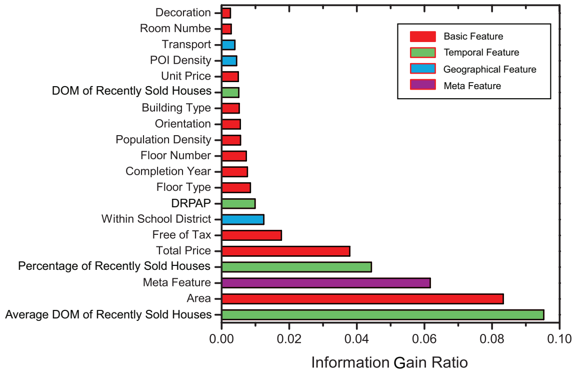


Figure 7: The top 20 features ranked by their information gain ratio.

We further examine the contributions of each individual feature. Specifically, we leverage the widely-used information gain ratio (IGR) as metric to determine which of the features are the most important. Figure 7 demonstrates the top 20 features ranked by their IGR (i.e., higher IGR indicates greater importance). In particular, for each feature that can be parameterized with different time span to generate different features (e.g., "Recent Sold DOM Average"), we only show the value of the most important one. From these results, we can find that the top 5 features that are most correlated with DOM are "Average DOM of Recent Sold", "Area", "Meta Feature", " Percentage of Recent Sold", and "Total Price" in which there are two temporal features, two raw attributes of houses, and the meta feature extracted using random forest. Interestingly, we find that the intuitive feature "Unit Price" has limited contribution to DOM prediction, which is because that most property listings within the same urban district have very similar unit price.

5.4 Evaluation of Different Regularizers

Here, we evaluate the effectiveness of different regularizers in our multi-task learning approach. To be specific, we compare our approach MLR-DOM with $L_{2,1}$ and L_1 regularizers, which are frequently used in multi-task learning. The results are shown in Table 7. We could observe that MLR-DOM has the best prediction performance in terms of MAE on all data sets, and MLR-DOM is always better than

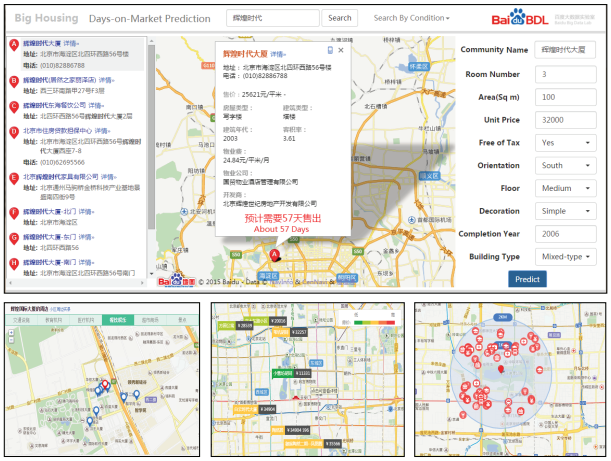
Table 7: The Performance of different regularizers.

Data Set	Method	rMSE	nMSE	MAE
D#1	L_1	60.7496	1.2701	49.2641
	$L_{2,1}$	**54.7373**	**1.0312**	38.8330
	MLR-DOM	54.7480	1.0316	**36.4876**
D#2	L_1	52.2400	1.2056	35.9265
	$L_{2,1}$	48.9935	1.0604	39.3837
	MLR-DOM	**46.3684**	**0.9498**	**32.7723**
D#3	L_1	41.3422	1.0966	31.3027
	$L_{2,1}$	39.3669	0.9943	31.7593
	MLR-DOM	**38.5147**	**0.9517**	**30.7996**

$L_{2,1}$ on each evaluation metric. Furthermore, with regard to rMSE and nMSE metrics, MLR-DOM always outperforms L_1 on all data sets, while MLR-DOM outperforms $L_{2,1}$ and have comparable performance with L_1 on D#2. Based on the above analysis, choosing graph regularizer and L_1 regularizer is reasonable for our multi-task learning approach.

5.5 The Prototype System

We have implemented a prototype system for DOM prediction using bootstrap (front-end framework for web development), angularJS (JavaScript MVW framework), and Django (a web framework in Python) along with MySQL. Specifically, when the user searches a community on the map, the system will show its position and detailed information, such as completion year and plot ratio. It also allows users to input profile information of their house, such as area, price, and then predict the DOM of the house given that information. If the user has not provided complete information, default average values will be used. We regularly update the database with new transaction data and train a new model upon each data refresh. The prediction model is trained offline so that users can get an instant result after submitting house profile information. Furthermore, the system also allows the visualization of real-estate data, such as visualizing the distribution of house prices in Beijing. Some screenshots of our demo system is shown in Figure 8.



辉煌时代大厦 详情»
地址：北京市海淀区北四环西路56号楼
电话：(010)82886788
售价：25621元/平米 -
房屋类型：写字楼
建筑类型：塔楼
建筑年代：2003
容积率：3.61
物业费：24.84元/平米/月
物业公司：国贸物业酒店管理有限公司
开发商：北京辉煌世纪房地产开发有限公司
预计需要57天售出
About 57 Days

Community Name: 辉煌时代大厦
Room Number: 3
Area(Sq m): 100
Unit Price: 32000
Free of Tax: Yes
Orientation: South
Floor: Medium
Decoration: Simple
Completion Year: 2006
Building Type: Mixed-type

Figure 8: The screenshots of our prototype system for DOM prediction.

6. RELATED WORK

In general, the related work can be grouped into two categories. The first category includes the studies of real estate DOM analysis and the second category includes the studies of multi-task learning.

6.1 Real Estate DOM Analysis

Analyzing the liquidity and popularity of markets is always important for different business sectors [32, 30, 27, 18, 6, 11]. In this paper, we focus on DOM of real estate, which measures the liquidity of real estate markets and shows the level of risk associated with real estate investments. There are a number of studies focused on analyzing the relationship between DOM and prices (both listed prices and sale prices). While there are studies to show positive relationships between sale prices and DOM [27, 18], some other studies claimed that the relationship between two variables is not significant [6]. In this paper, our focus is not on the relationship between these two variables. Instead, we focus on providing a solution to accurately predict DOM with extensive contextual information, such as house profile information and geo-social information.

Specifically, there are three types of methods for the DOM analysis. The first method is Ordinary Least Squares (OLS), which has the ability to test and make corrections for self-selection issues and is equally flexible in dealing with endogeneity issues, which are two major issues with the DOM analysis. However, non-normal error terms can lead to biased OLS coefficient estimates. Second, hazard models, often assuming a Weibull distribution of property marketing time, offer highly flexible functional specifications. However, there are the aforementioned self-selection and endogeneity problems with these models. Finally, instrumental variables models, usually in the form of 2SLS, allow for the joint estimation of simultaneously determined property price and property selling time. Nonetheless, there are criticisms related to the non-normality of the error term and the difficulty in calculating required inverse Mills ratios (IMR) to control for self-selection issues between variables of interest and property marketing time.

In this paper, our work is based on OLS and aims to improve its performance using machine learning techniques, such as feature engineering and multi-task learning. For OLS, there are two variants. For the first type of OLS, feature values (independent variables) and target variables appear in the model without non-linear transformation [28]. This is different from second type of OLS, where sale price, DOM, and some other variables will take logarithm before regression [22]. However, the second type of OLS usually performs very poor on MSE or MAE metrics, since the regression target is the logarithm of DOM rather than DOM. Therefore, we choose the first type of OLS as our baseline in this paper.

6.2 Multi-task Learning

Multi-task learning is a well-known machine learning methods to improve classification and regression performances by utilizing cross-task information. It first appears in the context of neural networks [10, 9]. Later, regularization-based multi-task learning starts to appear and forms an important research area [15, 14, 5].

Usually, regularization-based multi-task learning methods share the same framework, but differ in the choice of regularization terms used to represent assumptions of different types of task relationships. There are works which assume that all tasks are related and share a low-dimensional representation across a set of multiple related tasks. These works often select or learn a common set of shared features among the tasks [14, 5, 21]. Also, Obozinski and Argyriou *et al.* proposed joint feature learning for multi-task methods [5, 23]. Argyriou *et al.* generalized the well-known Lasso from single task case to multiple task case. [14]. In addition, some researchers believe that the assumption that all tasks are related is too strong and may not hold in real world. Therefore, they proposed other task relationship structures, such as clustered structures [17], tree structures [20], and graph structures [19, 13].

Finally, multi-task learning methods have been used in various fields, such as education, disease control, and computer vision. For instance, Argyriou has exploited multi-task learning for predicting students' exam scores. Bickel *et al.* has employed multi-task learning for HIV therapy screening [7] and Zhou *et al.* have proposed a multi-task formulation for predicting disease progression [31]. At last, Wang *et al.* have explored boosted multitask learning for web image and video search [29].

7. CONCLUSION

In this paper, we developed a comprehensive approach for measuring the liquidity of real estate markets. This approach provides a critical capacity for predicting the DOM of a given property listing, and thus enhances the information transparency between sellers and buyers. Specifically, we first investigated various contextual features for identifying the key factors that can affect the sale of real estates. Then, we developed a multi-task learning based regression approach for DOM prediction, which can effectively learn district-aware models for different property listings by integrating extracted contextual features. Finally, we presented experimental results to demonstrate the performance of our method with a large amount of real-world real estate data, and designed a prototype system showing the practical use of the liquidity analysis for real estate markets.

Acknowledgments

This work was partially supported by grants from National Science Foundation for Distinguished Young Scholars of China (Grant No. 61325010), the Natural Science Foundation of China (NSFC, Grant No.s 71329201, 61403358, 61572032, 71571093), and the Youth Innovation Promotion Association of CAS.

8. REFERENCES

[1] http://en.wikipedia.org/wiki/real_estate_investing.

[2] https://nlp.baidu.com.

[3] https://www.realtor.com.

[4] http://wiki.china.org.cn/wiki/index.php/five_policies_and_measures_to_regulate_real_estate_market.

[5] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. In *Machine Learning*, 73(3):243–272, 2008.

[6] J. D. Benefield and W. G. Hardin III. Does time-on-market measurement matter? In *The Journal of Real Estate Finance and Economics*, pages 1–22, 2013.

[7] S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer. Multi-task learning for hiv therapy screening. In ICML'2008, pages 56–63. ACM, 2008.

[8] K. Björklund, J. Alex Dadzie, and M. Wilhelmsson. Offer price, transaction price and time-on-market. In *Property Management*, 24(4):415–426, 2006.

[9] R. Caruana. Multitask learning. In *Machine learning*, 28(1):41–75, 1997.

[10] R. Caruna. Multitask learning: A knowledge-based source of inductive bias. In *ICML'1993*, pages 41–48, 1993.

[11] B. Chang, H. Zhu, Y. Ge, E. Chen, H. Xiong, and C. Tan. Predicting the popularity of online serials with autoregressive models. In *CIKM'2014*, pages 1339–1348. ACM, 2014.

[12] J. Chen, J. Zhou, and J. Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *KDD'2011*, pages 42–50. ACM, 2011.

[13] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, E. P. Xing, et al. Smoothing proximal gradient method for general structured sparse regression. In *The Annals of Applied Statistics*, 6(2):719–752, 2012.

[14] A. Evgeniou and M. Pontil. Multi-task feature learning. In *NIPS'2007*, 19:41, 2007.

[15] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *KDD'2004*, pages 109–117. ACM, 2004.

[16] Y. Fu, G. Liu, S. Papadimitriou, H. Xiong, Y. Ge, H. Zhu, and C. Zhu. Real estate ranking via mixed land-use latent models. In *KDD'2015*, ACM, pages 299–308, 2015.

[17] L. Jacob, J.-p. Vert, and F. R. Bach. Clustered multi-task learning: A convex formulation. In *NIPS'2009*, pages 745–752, 2009.

[18] R. Kalra, K. C. Chan, and P. Lai. Time on market and sales price of residential housing: A note. In *Journal of Economics and Finance*, 21(2):63–66, 1997.

[19] S. Kim and E. P. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. In *PLoS genetics*, 5(8):e1000587, 2009.

[20] S. Kim and E. P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML'2010*, pages 543–550, 2010.

[21] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *UAI'2009*, pages 339–348. AUAI Press, 2009.

[22] H. J. Munneke and A. Yavas. Incentives and performance in real estate brokerage. In *The Journal of Real Estate Finance and Economics*, 22(1):5–21, 2001.

[23] G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. In *Statistics and Computing*, 20(2):231–252, 2010.

[24] M. R. Segal. Machine learning benchmarks and random forest regression. In *Center for Bioinformatics & Molecular Biostatistics*, 2004.

[25] C. R. Taylor. Time-on-the-market as a sign of quality. In *The Review of Economic Studies*, 66(3):555–578, 1999.

[26] W. R. Tobler. A computer movie simulating urban growth in the detroit region. In *Economic geography*, pages 234–240, 1970.

[27] R. R. Trippi. Estimating the relationship between price and time to sale for investment property. In *Management Science*, 23(8):838–842, 1977.

[28] C. Tucker, J. Zhang, and T. Zhu. Days on market and home sales. In *The RAND Journal of Economics*, 44(2):337–360, 2013.

[29] X. Wang, C. Zhang, and Z. Zhang. Boosted multi-task learning for face verification with applications to web image and video search. In *CVPR'2009*, pages 142–149. IEEE, 2009.

[30] L. Wu, Q. Liu, E. Chen, X. Xie, and C. Tan. Product adoption rate prediction: A multi-factor view. In *SDM'2015*, pages 154–162, 2015.

[31] J. Zhou, L. Yuan, J. Liu, and J. Ye. A multi-task learning formulation for predicting disease progression. In *KDD'2011*, pages 814–822. ACM, 2011.

[32] H. Zhu, C. Liu, Y. Ge, H. Xiong, and E. Chen. Popularity modeling for mobile apps: A sequential approach. *IEEE Trans. Cybernetics*, 45(7):1303–1314, 2015.