Prediction of Days-On-Market for Single-Family Homes

K. Galbraithy[1], R. R. Hashemi[2], O. M. Ardakani[3], J. S. Beck[4]

[1, 2] Department of Computer Science, Georgia Southern University, Statesboro, GA, USA, kg06220@georgiasouthern.edu, and rayhashemi@georgiasouthern.edu

[3, 4] Department of Economics, Georgia Southern University, Savannah, GA, USA, oardakani@georgiasouthern.edu, and jbeck@georgiasouthern.edu

Abstract — The number of days that a home stays on the housing market (Days-On-Market—DOM) provides crucial information at both micro-level (behavior associated with the buyer's/seller's decision) and macro-level (risk associated with real estate investments and also the housing bubbles' identification). Housing data has a mixture of simple and complex attributes. Simple attributes have one value, whereas complex attributes have a variable-length array of values per property and, therefore, their inclusion in the prediction of DOM is a major challenge. The goal of this research effort is: (a) Providing for complex attributes in DOM's prediction, (b) Analyzing, designing, and implementing a DOM prediction's package using Linear Regression and Naïve Bayesian, separately, and (c) Establishing the superiority and robustness of the underline models.

Keywords: Association Analysis, Linear Regression, Naïve Bayesian, Discretization of Multi-valued Attributes, and Housing Market Analysis

Introduction

A single-family home listed in a housing market has a large set of attributes that potential buyers consider when deciding on the purchase of a home. One such attribute is the Days-On-Market (DOM), which represents the number of days a home stays on the market. DOM is an important indicator that provides information about the behavior of the real estate market. At the micro-level, DOM affects the buyer's and seller's decisions [1][2][3]. Buyers consider DOM as an indication of the home's condition and the seller's motivation for selling the house. Sellers use the average DOM for comparative homes as one strong factor in pricing their properties.

At the macro level, DOM is a measure of liquidity in the housing market and indicates the risk associated with real estate investments. It also helps policymakers to identify the housing bubbles [4][5].

There is a relatively small number of investigations on the prediction of DOM for a given housing market. Statistical approaches [6][7][8] and, more recently, machine learning approaches [9][10] are the methodologies employed in the prediction of DOM.

One of the major challenges in the prediction of the DOM is having a mixture of *simple* and *complex* attributes. A simple attribute carries one value for a given property. The *number of bedrooms*, *list price*, *number of garages* are among the simple attributes.

A complex attribute may have an array of values for a given property. For example, the attribute of *floor description* has an array of values for a property because a property may have more than one type of covered floor. A complex attribute borrows its array of values from a domain that lists possible values of the attribute for all properties. As a result, the array's size and array's values may differ from one property to the next. The *exterior*, *construction type*, *community*, *custom rooms* are among the complex attributes. To the best of our knowledge, handling the complex attributes in prediction of DOM has not been reported in literature.

The goal of this research effort is tri-fold: (a) inclusion of complex attributes in DOM's prediction for single-family homes, (b) Analysis, design, and implementation of two prediction models of Linear Regression and Naïve Bayesian to predict DOM, and (c) Comparing the results to establish the prediction superiority and robustness of the models. To meet the goal, the following objectives are completed (i) Cleaning the single-family homes' dataset and discretizing data (ii) Identifying the relevant attributes to DOM, (iii) Developing the two Prediction models of Linear Regression and Naïve Bayesian and comparing their performances, and (iv) Examining the robustness of the prediction models.

The rest of the paper is organized as follows. The Methodology is presented in Section 2. The Empirical Results are discussed in Section 3. The Conclusions and Future Research are covered in Section 4.

Methodology

The methodologies for cleaning and discretizing data, identifying the relevant attributes to DOM, and developing the two prediction models of Linear Regression and Naïve Bayesian are covered in the following three sub-sections.

Cleaning and Discretization of Data

The original dataset consists of 520 attributes and 34,730 data records for the real estate properties of several cities. And for duration of 9 years (2007-2016). Since we are interested in the DOM investigation for the housing market of one specific city, records that belong to the housing market of other cities are removed. We have also a small number of records with missing data (<0.24%) that are removed.

The dataset includes a number of disaggregated attributes that are *aggregated* as one attribute. For example, all disaggregated attributes accommodating the components of a property's Zip-code are aggregated under one attribute, Zip. There are also *after-the-fact* attributes that carry data about the completion of the selling process. These attributes are removed. The *Closing-Date*, *Closing-Cost*, *Sold-Price* are among this type of attributes. In addition, there are *secondary* attributes with no effects on DOM and they are removed. The *seller and buyer agents' appointments log* and the *listing agency address* are samples of the secondary attributes. Furthermore, the cleaning process creates *trivial* attributes that are dismissed. For example, having an attribute for the city name in which the real estate property is resided becomes a trivial attribute, when the entire dataset is about the sale of single-family homes of one specific city. The outcome of the cleaning process is a dataset with 50 attributes and 20,866 records.

To discretize data, attributes are divided into two major groups of *simple* and *complex*. A simple attribute has only one value for each real estate property. Simple attributes are further divided into *category* and *continuous,* based on the nature of their values. For example, attributes of the *number-of-bedroom*, *number-of-garage*, and *number-of-bath* are among the category attributes. For these attributes, we used their categorical data as their discrete values. In the case, that a category attribute has linguistic values, an integer number is assigned to each linguistic value and the assigned integer numbers are used as the discrete values for the attribute. However, there are category attributes with too many categorical data and, hence, too many discrete values. For these category attributes some of the categorical data are collapsed to bring down the number of discrete values. The attribute of the number-of-room is in this group. The number-of-rooms may go up to 12 and so the number of discrete values. To reduce the number of discrete values for the number-of rooms, more than six rooms are collapsed into one category.

Each continuous attribute is discretized either by an equal-width histogram method or an equal-size histogram method [11]. In the former method, the difference between the minimum and maximum continuous values for the attribute is divided into n equal width. That is, properties are divided into n groups and the discrete values of 1 to n are assigned to the continuous attribute accordingly. In the later method, the number of real estate properties for each group remains the same and the width changes.

A complex attribute has an array of values for each property and these values are categorical data. The array's size and array's values could be different from one property to the next. To discretize these attributes, the following three steps are taken:

1. We consider each unique categorical value as *1-value itemset* and count the frequency of its appearance in the attribute for the entire dataset. Those 1-value

itemsets with count less than a threshold (*support count—SC*) are dismissed as disqualified itemsets. We build all the possible *2-value itemsets* out of the qualified 1-value itemsets by creating all of their possible pairs. The qualified 2-value itemsets are identified by counting their frequency in the attribute and dismissing those with counts less than SC. In the next iteration *3-value itemsets* are generated and disqualified ones are dismissed.

2. We continue this process until no new qualified itemset could be generated. Suppose in the iteration number k+1we have reached to this point. This means that at the k^{th} iteration we have *k-value itemsets* that are qualified (k is referred to as the *level* of the itemsets). Therefore, we have qualified itemsets from level 1 to level k. Creation of the itemsets are captured by the algorithm One which is a version of the Apriori Algorithm and shown in Fig. 1. [12].

```
Algorithm ONE
Input:     Complex attribute of C and a support_count (SC).
Output:    K-value itemset(s)
Method:
  i = 1;
  go = 1;
  Build all the possible 1-value itemsets;
  Repeat (go)
          For each i-value itemset do:
               Count the frequency of i-value itemset;
               If (count  ≥ SC)
                   Then Keep the i-value itemset as a qualified itemset;
                   Else dismiss the i-value itemset;
          End;
          i++;
          Build all the possible i-value itemsets out of the qualified (i-
          1)-value itemsets;
          If (no i-value itemsets can be built)
             Then go = 0;
  End;
End;
```

Fig. 1 A Version of Apriori Algorithm

3. Starting from level j=1 and remove all the itemsets in level j if they are a subset of itemsets in level j+1. Assign a discrete value of 1 to k to all survived qualified itemsets in levels of 1 to k, respectively. This means that if we have more than one k-value itemsets in level k, for example, then all of them have assigned the discrete value of k. The same is true for all the levels. One may ask, why each itemset in level k, for example, is not assigned a different discrete value. The answer lies on

the fact that each itemset in level k has the same attraction to different group of buyers.

4. If for complex attribute X, for example, there is a real estate property that its X-value does not match any discrete values produced by the above steps, then the discrete value is zero. Such a case is rarely possible, if the X-value for the property is made up of all disqualified 1-value itemsets.

The decision attribute of DOM is discretized into two values of 1 and 2. DOM = 1 (Short) means that the property is maximum of 180 days (six months) in the market. DOM = 2 (Long) means that the property is more than 6 months in the market.

Identification of the Relevant Attributes

An attribute is considered relevant if it influences the DOM attribute. As the first step, we used the Cramer's V Measure of Association [13] to identify those attributes who have high correlations with other attributes (> 0.7), and they are removed. For further identification of the relevant attributes, we use two different methods of Gain Measure of Association [14] and Ratio Measure of Association [15]. In both methods, DOM is considered as the function of the rest of the attributes. A short description of the Cramer's V Measure of Association, Gain Measure of Association, and Ratio Measure of Association are as follows:

Cramer's V Measure of Association Delivers the correlations among the attributes of the cleaned dataset. Cramer's V measure of association is suitable when discrete data are used. The measure is based on the chi-squared test of association. Let χ^2 be Pearson's chi-squared statistics. For an m × n contingency table, the Cramer's V statistic can be written as Formula (1)

$$V = \sqrt{\frac{x^2}{N[\min(m,n)-1]}} \qquad (1)$$

Where, N is the number of records and V ranges between 0 and 1, where 1 indicates the highest correlation.

Gain Measure of Association Deliveres a measure of gain for DOM. The attributes for which DOM has a higher gain are considered more relevant to DOM. Gain is measured by applying:

a. Formula (2) to calculate the entropy for a given attribute, X, in the dataset.

$$E(X) = -p_1 {*} \log_2 p_1 - \ldots - p_k {*} \log_2 p_k \qquad (2)$$

Where, k is the number of the discrete values in the attribute and p_i is probability for discrete value of i in the attribute X (i.e., the frequency of discrete value i in the attribute to the total number of records in the dataset).

b. Formula (3) to calculate the branching measures of entropy for different discrete values of attribute X.

$$B(X) = \sum_{i=1}^{n} p_{v_i} * E(d_{v_i}) \quad (3)$$

Where, n is the number of discrete values in X, v_i is a discrete value in X, p_{vi} is the probability of the discrete value of v_i in X, d_{v_i} is a subset of the records in the dataset for which the value of X is v_i, and $E(d_{v_i})$is the entropy of the DOM in the dataset d_{v_i}.

c. Formula (4) to calculate the measure of gain, G, for DOM in reference to a given attribute X.

$$Gx(DOM) = E(DOM)\text{-}B(X) \quad (4)$$

Ratio Measure of Association Delivers a measure of Ratio for DOM. As a matter of fact, the Gain Measure is generally higher for those attributes with a higher number of discrete values. To eliminate this sensitivity, the Ratio measure is used. In reference to a given attribute X, the ratio measure, Rx, of DOM is calculated by Formula (5).

$$Rx(DOM) = Gx(DOM)/\ N \quad (5)$$

Where, N is the number of discrete values for X and Gx(DOM) is the gain measure of DOM for the attribute X.

Development of the Prediction models

Two prediction models of Linear Regression [15] and Naïve Bayesian [16] are used separately to predict the DOM. Two prediction models are used not only to compare the prediction power of the two models but also their robustness.

Linear Regression: The prediction model can be written as formula (6):

$$DOMt = Xt\beta + \varepsilon_t \quad (6)$$

Where, DOM_t is the N*1 days-on-market vector, X_t is the N*k matrix of housing attributes, β is a vector of parameters indicating the relationship between DOM and the rest of the attribute test, ε_t is the error term. The least-squares estimator is obtained by minimizing the sum of a quadratic loss function, formula (7):

$$L(DOM, X\beta) = (DOMt - Xt\beta)^2 \quad (7)$$

This solution provides the coefficients needed for predicting DOM. We examine its performance using an out-of-sample forecasting scheme. The Linear Regression is executed using the R statistical package [17]. The model learns from a training set and is tested against a test set. For every test record, the predicted value for DOM is compared to the actual DOM value of the test record. (This is possible because the test set contained a number of randomly chosen records from our cleaned and discretized real estate dataset). If the predicted value is within a certain margin of error, then the prediction is considered accurate.

Naïve Bayesian system: This prediction model delivers a predicted DOM value for a test record, r, using a training set. The test record is made-up of K attributes of $A_1, \ldots, A_k$ and each record of the training set is made-up of (K+1) attributes. The first K attributes are the same as the K attributes of the test record and the extra attribute is the DOM attribute. For the discrete value of j (j = 1 to m) in DOM, we calculate the conditional probability of the test record having the discrete value of j using formula (8). Consequently, m different conditional probabilities are calculated for the test record r. The discrete value with the highest probability is the winner and it is selected as the predicted DOM value for r.

$$P(DOM = j \,/\, r) = \frac{P(DOM=j) * \prod_i^k P(A_i = v \,|DOM=j)}{P(r)} \qquad (8)$$

Where, A_i is the i-th attribute of the record r and v is its value, k is the number of attributes in the test record of r, j is a discrete value for DOM attribute, P(DOM = j) is the probability of DOM with value of j in the training set, and P(A_i = v |DOM = j) is the conditional probability of attribute Ai with value of v in the training set, given DOM = j. P(r) does not need to be calculated because it will have the same value for all of the P(DOM= j | r) (for j =1 to m).

In the case of a tie between the conditional probabilities, a failure in prediction of DOM, for r, has been reached. The prediction accuracy could be easily determined because r borrowed from the original dataset.

The robustness measure of a predictive model refers to a measure of how effective the model is when encountering a noisy dataset [18]. We selected the version of our dataset prior to removal of the irrelevant attributes, using Gain Measure of Association, as our noisy dataset. The degradation of the predictive power of the two models due to the presence of noise to the dataset are measured by Formula (9).

$$Degradation = PC/PN \qquad (9)$$

Where, PC is the percentage of drop in the prediction accuracy and PN is the percentage of added noise to the dataset.

Empirical Results

After cleaning and discretizing the dataset, the Cramer's V Measure of Association identified three attributes of *covenant fee*, *recreation-facilities*, and *Neighborhood-Highschool* attributes that are highly correlated with other attributes and they are removed from the dataset.

The identification of the relevant attributes is completed using the two methods of Gain and Ratio measure of associations separately. The findings are shared and consulted with two real estate agents as experts. The most relevant attributes identified by the Ratio measures of association are selected to work with. As a result, our working real estate dataset has 20,866 records and 18 attributes.

The application of the Linear Regression on 100 randomly chosen pairs of training (80% of the records) and test (20% of the records) sets are completed for four different error's margins of 0.4, 0.35, 0.3, 0.25, 0.20, and 0.15. The averages for the accuracy of the prediction of the test records for 100 random pairs of training and test sets are shown in Table 1. The detailed predictions of for short-period (DOM=1) and long-period (DOM=2) for the 0.4 error margin is shown in Table 2.a.

The Naïve Bayesian prediction model also applied on 100 randomly chosen pairs of training and test sets (with 80% of the records in the training set and 20% of the records in the test set). The average results are shown in Table 2.b. The choice of the margin of error is not directly applicable to the Naïve Bayesian .

Table 1 The average for the correct predictions of the test records of 100 random pairs of training (80% of the records) and test (20% of the records) sets by the Linear Regression.

	Margin of Error					
	0.40	0.35	0.30	0.25	0.20	0.15
Mean	0.96	0.94	0.93	0.91	0.88	0.76
Accuracy	95%	94%v	93%	91%	88%	76%

Table 2 The average for the correct predictions of the test records of 100 random pairs of training (80% of the records) and test (20% of the records) sets by: (a) the Linear Regression, and (b) the Naïve Bayesian .

Predicted DOM Value	Actual DOM Values		
	1	2	Total
1	3230	143	3373
2	69	647	716
Total	3299	790	4089

(a)

Predicted DOM Values	Actual DOM Values		
	1	2	Total
1	2957	649	3606
2	333	152	485
Total	3290	801	4091

(b)

Table 3 Different Statistics for the Linear Regression and model and the Naïve Bayesian.

Statistics	Linear Regression Model	Naïve Bayesian Model
% of true prediction of the short-period (DOM=1)	3230/3299 =98%	2957/3290=90%
% of false prediction of the short-period (DOM=1)	69/3299 =2%	333/3290=10%
% of true prediction of the long-period (DOM=2)	647/790=82%	649/801=81%
% of false prediction of the long-period (DOM=2)	143/790=18%	152/801=19%
Error rate	212/4089=5%	982/4091=24%
Accuracy	3877/4089=95%	3109/4091=76%

Different statistics are calculated for both the Linear Regression and the Naïve Bayesian models and are shown in Table 3.

To compare the robustness of the two prediction models in reference to our real estate properties, we added 27 attributes from the list of irrelevant attributes as noise to our dataset. This addition increased the number of attributes in our dataset from 18 to 45. In fact, we increased the number of attributes by more than 150% to introduce heavy noise to our dataset. The amount of noise is the predominant in this noisy dataset.

Both predictive models are applied on the noisy dataset for comparing their robustness. The average for the prediction accuracy of the test records of 100 random pairs of training (80% of the records) and test (20% of the records) sets by the Linear Regression and Naïve Bayesian models along with their prediction power degradation are depicted in Table 4 using formula (9). Based on the results in Table 4 the degradation in the prediction power of the Naïve Bayesian model is as twice as the degradation for the Linear Regression.

Table 4 The average degradation of the prediction power for the Linear Regression and Naïve Bayesian models using noisy dataset

Model	Accuracy Prediction		% of Degradation of Prediction Power
	Cleaned Data	Noisy Data	
Linear regression	95%	87%	8/150=0.5%
Naïve Bayesian Model	76%	60%	16/150=1%

Conclusions, and Future Research

The results revealed the superiority of the Linear Regression performance over the Naïve Bayesian model for the margin of error of greater than or equal to 0.2 (Reported in Table 3). However, for the margin of error less than 0.2, the performance of the Linear Regression model came extremely close to the performance of the Naïve Bayesian model. In reference to the Naïve Bayesian model, one could argue that for a given test record, the conditional probability of DOM=1 and DOM =2 may differ in their, say, fifth digit. This means that the margin of error is a minimum of 0.00001. The maximum difference between the two conditional probabilities is 1, the minimum difference is close to zero, and on average, the difference is 0.5. If this margin of error is used for the Linear Regression model, the prediction accuracy of DOM is 100%. Therefore, Linear Regression has a higher power of prediction. The robustness testing of the two models also supported the fact that Linear Regression is more robust.

As future research, we will (a) increase the discrete values for DOM to narrow down its prediction and (b) model the causality of "days-on-market" (DOM) for single-family homes. This modeling will be used to investigate causes behind DOM that could be manipulated by a seller/agent for a desirable outcome.

References

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

[2] Knight, J. R. (2002). Listing price, time on market, and ultimate selling price: Causes and effects of listing price changes. Real Estate Economics, 30(2), 213–237.

[3] Benefield, J. D., & Hardin, W. G., III. (2015). Does time-on-market measurement matter? The Journal of Real Estate Finance and Economics, 50, 52–73.

[4] Zhu H., Xiong H., Tang F., Liu Q. , Ge Y., Chen E., & Fu Y. (2016). Days on Market: Measuring Liquidity in Real Estate Markets. , Proceedings of KDD '16, San Francisco, CA, USA.

[5] Bourassa A. C., Hoesli M., Oikarinen E. (2019). Measuring House Price Bubbles. Real Estate Economics, 47(2), 534-563.

[6] Jud G. D. (1996). Time on the market: the impact of residential brokerage. Journal of Real Estate Research, 12(3), 447–458.

[7] Belkin D. D. (1976). An empirical study of time on market using multidimensional segmentation of housing markets. Real Estate Economics, 4(2)57-75.

[8] Miller, N. (1978). Time on the market and selling price. Real Estate Economics, 6(2), 164–174.

[9] Castelli, M., Dobreva, M., Henriques, R., & Vanneschi, L. (2020). Predicting days on market to optimize real estate sales, strategy. Complexity, (3)1–22.

[10] Ermolin S. V. (date visited: August 2020). Predicting Days-On-Market for Residential Real Estate Sales. https://pdfs.semanticscholar.org/8dd1/6fc494f30a0a33e521243473966a26350025.pdf

[11] Han J., Kamber M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Publisher: Morgan Kaufmann.

[12] Hashemi R., Le Blanc L., Bahrami A., Bahar M., and Traywick B. (2009). Association Analysis of the Alumni Giving: A Formal Concept Analysis. the International Journal of Intelligent Information Technologies, 5(2), 17-32.

[13] Cramer, Harald. 1946. Mathematical Methods of Statistics. Princeton: Princeton University Press.

[14] Quinlan J. R. (1986). Introduction to Decision Trees. Journal of Machine Learning. Publisher: Kluwer Academic Publishers. 1(1), 81-106.

[15] Quinlan J. R. (1993). C4.5: Programs for Machine Learning, Publisher: Morgan Kaufman.

[16] Hashemi R., Ardakani O., Bahrami A., Young J., and Campbell R.(2018). A Mining Driven Decision Support System for Joining the European Monetary Union, The Eighth International Conference on Advances in Information Mining and Management (IMMM'18), Barcelona, Spain, pp. 39-45.

[17] The R Foundation (date visited: March 2020). The R Project For Statistical Computing https://cran.r-project.org/bin/macosx/

[18] Marco Avella-Medina (2020). The Role of Robust Statistics in Private Data Analysis, CHANCE, 33(4) 37-42.