Information Value of Property Description: A Machine Learning Approach *

Lily Shen
Clemson University

November 13, 2018

Abstract

This paper employs a ML–Hedonic approach to quantify the value of uniqueness, a type of "soft" information embedded in real estate advertisements. We first propose an unsupervised learning algorithm to quantify levels of semantic deviation ("uniqueness") in descriptions, the textual portions of real estate advertisements. We then estimated the impact of description uniqueness on real estate transaction outcomes using linear hedonic pricing models. The results indicate textual data disseminate information that numerical data cannot capture, and property descriptions effectively narrow the information gap between structured real estate data and the houses by conveying "soft" information about unique house features. A one standard deviation (0.08) increase in description uniqueness compared to neighboring properties leads to a 5.6% increase in property sale prices and a 2.3–day delay in the closing time, controlling for house characteristics, transaction circumstances, and agent unobservables. This paper provides theoretical and empirical insights on how to utilize the emerging Machine Learning tools in economic research.

*Lily Shen, yannans@clemson.edu, Clemson University Finance department. For helpful comments and discussions we would like to thank seminar attendees at Clemson University/Finance, the Federal Reserve Bank of Cleveland, and the Federal Reserve Bank of Atlanta. We are especially grateful for comments and suggestions from Brent Ambrose, James Conklin,Chris Cunningham, Bruce Fallick, Kris Gerardi, Roberto Pinheiro, Stuart Rosenthal, Vincent Intintoli, and Stephen Ross. All errors are our own.

1 Introduction

Real estate is an essential component of the U.S economy. In 2015, the real estate industry generated $3 trillion of revenue, which accounted for 17.3% of GDP.[1] Although many efforts have been made by researchers and practitioners aiming to model real estate prices, one obstacle that has not been overcome is how to systematically identify and control for house characteristics that are not reported by the existing structured real estate databases, commonly known as "soft" information (Liberti and Petersen, 2018). For example, while the number of bedrooms is a type of "hard" house information, the overall uniqueness of a house compared to its cohorts is an unreported "soft" feature. According to a 2018 Wall Street Journal article, institutional investors who buy and sell hundreds of houses on a daily basis utilize the newly advanced Artificial Intelligence technologies to extract "soft" information from unstructured real estate data, such as the textual house descriptions.[2]

Rosen (1974) provides a theoretical framework enabling the sale price of a house to be modeled with a hedonic function of its physical and location attributes. Following this seminal paper, a strand of literature has been developed using the numerical portion of real estate transactions data to assess the price impacts for certain housing characteristics and neighborhood disamenities (Palmquist, 1984, Lindenthal, 2017, Muehlenbachs et al., 2015, Bernstein et al., 2018).[3] Although Liberti and Petersen (2018) and Garmaise and Moskowitz (2004) point out that houses are heterogeneous goods for which some characteristics cannot be easily captured by numerical data ("hard" information), the widely used hedonic framework does not provide a solution to control for unobserved heterogeneities that cannot be easily transmitted in impersonal ways ("soft" information).

A few recent studies have recognized the information potential of textual data in shedding light on different aspects of the business world. Tetlock (2007), Garcia (2013), and Loughran

[1] 2016 National Association of Realtors report: Economic Impact of Real Estate Activity

[2] Dezember, Ryan, "How to Buy a House the Wall Street Way" The Wall Street Journal, September 16, 2018.

[3] Examples of numerical data including but not limited to asking price, sale price, size, structure age, property type, location, and market condition

and McDonald (2011) find words that convey positive/negative meanings used in the popular press articles and 10-K reports can explain positive/negative stock returns. The evidence on the price impacts of frequent keywords in real estate advertisements thus far has been mixed. On one hand, Levitt and Syverson (2008), Rutherford and Yavas (2005), and Nowak and Smith (2017) find the inclusion of indicator variables for positive/negative words and short phrases in real estate advertisements can reduce omitted variable biases; On the other hand, Goodwin (2014) and Pryce (2008) point out the effects of positive/negative words on real estate prices are not consistent across different word classes.

Our study extends the keyword-based textual analysis literature by focusing on the semantic deviations between real estate advertisements. Property descriptions are the written portion of a real estate advertisement that summarizes critical features of the underlying house. We combine a novel Machine Learning method with hedonic pricing methods (hereafter referred to as the "ML–Hedonic approach") to quantify the price impact of product uniqueness, a type of "soft" information, in the real estate market. First, we train our ML algorithm to understand the semantic meaning of real estate descriptions, allowing us to quantify the uniqueness of a house in a neighborhood. Next, we estimate the impact of description uniqueness on real estate transaction outcomes using a linear hedonic model.

It is important to emphasize that the Machine Learning method used in this study is unsupervised, which is very different from the supervised algorithms discussed in Mullainathan and Spiess (2017). Supervised Machine Learning methods are often used to generate out–of–sample predictions based on a large number of training data.[4] In other words, the goal of supervised learning is to produce an inferred function to map input variables to desired output values based on human–labeled data. In the example illustrated in Mullainathan and Spiess (2017), the output variable is real estate sale prices and the input variables are numerical house characteristics. Unsupervised learning does not require labeled outputs, and is only used to infer the natural structure of data.

[4]Accuracy of the prediction results are highly dependent on the number of training data.

Figure 1 illustrates the fundamental differences between the traditional ML approach using supervised learning methods and our ML–Hedonic approach. A drawback of using supervised learning to predict house prices directly is that the middle procedure between input variables and prediction results is a "black box." In this paper, unsupervised learning is only used to compute a numerical uniqueness measure of a house based on its textual description. The impact of description uniqueness on real estate sale prices is estimated in a classic hedonic model, controlling for house physical characteristics, transaction circumstances, and agent unobservables. The ML–Hedonic approach enables us to draw clear economic inferences about the relationship between description uniqueness and transaction outcomes.

Using a data set that encompass more than 40,000 single–family houses sold in Atlanta, GA from January 2010 to December 2017, the analysis results suggest houses advertised by unique descriptions are associated with higher sale prices than those advertised by less unique descriptions. Comparisons among houses located in geographical proximity show that a one standard deviation (0.08) increase in description uniqueness leads to a 5.6% increase in property sale prices. Most of the price premium is driven by unique features of houses while language uniqueness has minimal impacts on sale prices.

To our best knowledge, we are the first to measure real estate uniqueness using textual data. Haurin (1988) models real estate "atypicality" using the observable numerical house features and sale prices. Lindenthal (2017) compares architectural design similarity using satellite photos. In this paper, we focus on the unobserved house features using real estate advertisements.

More broadly, this study provides several theoretical and empirical insights in response to the newly available ML tools in economics research. First, our ML algorithm defines the meanings of words within their contexts, addressing criticism for the keyword–based studies raised by Larcker and Zakolyukina (2012): "simply counting words (bag–of–words) ignores important context and background knowledge." Second, the ML–Hedonic approach

used in this study provides an example of the integration of unsupervised learning methods into economic analysis. In summary, our context–based ML algorithm can be applied to extract information from a wide range of textual documents to study a variety of economic phenomena.

The rest of the paper is organized as follows. We discuss the unsupervised Machine Learning algorithm and the empirical hedonic model in Section 2. We present the descriptive statistics of our real estate data and the description uniqueness score variable in Section 3. We document the impact of description uniqueness on real estate sale prices and liquidity in Section 4, and in Section 5 we discuss the plausible link between market market experiences and the effectiveness of unique descriptions. Section 6 concludes the paper.

2 Methodology

The ML–Hedonic approach follows three steps. First, we train our Machine Learning semantic analysis algorithm to understand the semantic meaning of real estate descriptions. Each description is represented as a vector in a high–dimensional vector space based on its contents and the distance between two vectors represents the pairwise difference between two houses. Second, we calculate the average pairwise difference between every house i in our data and its neighboring houses to identify the uniqueness of house i. Finally, we estimate the impact of description uniqueness on real estate sale prices using a linear hedonic model. We introduce the ML model in Section 2.1–2.2, and in Section 2.3 we describe our hedonic specification.

2.1 The Machine Learning Semantic Analysis Model

Natural Language Processing (NLP) algorithms use mathematical and statistical methods to help the computer learn and process human language. Applications of NLP include language translation, speech recognition, automatic summarization, natural language understanding,

etc., whereas our study focuses on the last task.

In this study, we implement the paragraph vector (PV) method, a Neural Network approach to obtain vector representations of real estate descriptions. Dai et al. (2015) compared the PV method against other textual analysis algorithms, including the widely used Bag–of–Words method on the analysis of 4,490,000 Wikipedia articles and 886,000 technical research papers, and concluded the PV method strictly outperformed the other methods. Inspired by Dai et al. (2015), our Neural Network Machine Learning algorithm naturally preserves the meaning of textual documents within its context and therefore complements those keyword–based textual studies mentioned previously. This approach offers the following advantages in analyzing property descriptions data:

First, our algorithm is more suitable for detecting nuances of human language compared to sentiment analysis methods based on word polarity. False–positive words are often used to glorify negative features of houses in the descriptions. For example, "good," "potential," "cozy," "cute," and "original" are all positive words in daily uses. However, in real estate descriptions "good potential" is often used to describe houses that require extensive renovation, "cute" and "cozy" are used to describe small houses, and "original" is used to describe old houses.

Second, our algorithm is more suitable for understanding abbreviations and typos, compared to sentiment analysis methods based on counting word frequency. Unlike formal documents such as 10–K reports and newspaper articles that are well polished before being released to the public, typos are often found in property descriptions (e.g. "morgage" vs. "mortgage"). In addition, the MLS systems impose a 250–word limit on description length, and thus the full spelling of a word might be replaced with an unstandardized abbreviation to save space. For example, "tender love and care" is a common expression in descriptions to describe old houses that need renovation. Depending on space availability, it can be written as "tender loving care," "tender love care" or "TLC". Unstandardized abbreviations and typos would have been dropped by previous algorithms based on counting word frequency.

Since our algorithm defines the meaning of textual data within their contexts, it is able to understand that all four expressions have the same meaning.

The learning model used in this paper is unsupervised, which does not require any prior assumption or knowledge about house descriptions. The training goal of this ML model is to convert textual real estate descriptions into vectors (vectorization of textual data). We use a simple Neural Network model with three layers: the input layer, the hidden layer, and the output layer.

Figure 2 illustrates the process of projecting a real estate description into a 7–dimensional vector space. It is important to emphasize that this simplified example is only created for demonstration purposes. The actual learning algorithm is more sophisticated and projects real estate descriptions into a space with over 150 dimensions.

In this study, each input item (w_{In}) is the house description with certain words (w_{Out}) removed. We are training the Neural Network model to successfully understand which words match the context with the highest probability. The algorithm is built upon a fundamental linguistic principle: words used together often share syntactic and semantic relations with each other, commonly known as "You shall know a word by the company it keeps" (Firth, 1957).

The following example demonstrates how the algorithm defines the meaning of "Southern" by its contexts in the house descriptions:

- This home is a graceful Southern beauty with rare stately double–front porches.
- Southern elegance in the Georgian style renovated for today.
- Graceful Southern charm!
- Exquisite Southern living, backyard w/stunning granite pool.

Semantically and syntactically, "Southern" is related to "elegance," "exquisite," "beauty," "graceful," etc. The ML algorithm learns the possible uses of "Southern" by analyzing the contexts.

We train the Neural Network model iteratively to get a vector representation of each description. First, the input item w_{In} is projected into the n–dimensional space by a weighting matrix. We are specifically interested in constructing this weighting matrix for the corresponding input. Since a paragraph can be analyzed as a combination of sentences composed of words, we include a paragraph vector ("D" on Figure 2) to represent the overall paragraph weights between the input layer and the hidden layer. Next, we reverse the vectorization process to obtain a conditional probability, $p(w_{Out}|w_{In})$, to map the projected vectors back to the correct output w_{Out}. The probabilities $p(w_{Out}|w_{In})$ are evaluated iteratively using the Softmax function[5], a log–linear classification method, during every training cycle until convergence, as depicted in the following equation:

$$p(w_{Out}|w_{In}) = \frac{exp(\boldsymbol{v}_{Out}^T \boldsymbol{v}_{In})}{\sum_o exp(\boldsymbol{v}_o^T \boldsymbol{v}_{In})} \tag{1}$$

where v_{In} and v_{Out} are vector representations of the input house description (w_{In}) and corresponding output text (w_{Out}).

Iteratively, we maximize the probability of getting the correct outputs through fine–tuning of θ=$\{$ v_{In}, v_{Out} $\}$, while minimizing the loss function E:

$$\begin{aligned} \arg\max_{\theta} p\left(w_{Out}|w_{In};\theta\right) &= \arg\max_{\theta} \log\left(p\left(w_{Out}|w_{In}\right)\right) \\ &= \arg\max_{\theta} \left[\boldsymbol{v}_{Out}^T \boldsymbol{v}_{In} - \log \sum\nolimits_o \exp\left(\boldsymbol{v}_o^T \boldsymbol{v}_{In}\right)\right] \end{aligned} \tag{2}$$

Correspondingly, the training goal is to minimize the loss function below:

$$E = \log \sum\nolimits_i \exp\left(\boldsymbol{u}_o\right) - \boldsymbol{u}_{out}^* = \log \sum\nolimits_o \exp\left(\boldsymbol{v}_o^T \boldsymbol{v}_{In}\right) - \boldsymbol{v}_{Out}^T \boldsymbol{v}_{In} \tag{3}$$

where u_{out}^* is the output vector corresponding to the ground truth output words. We also define a generic output score, $u_o = v_o^T v_i$, to be the output probability based on a given input

[5]The Softmax function is a generalization of the logistic function to calculate categorical probability used in Artificial Neural Networks

word.

While minimizing the loss function (Equation 3), we can backtrack the vector representation of each individual input through a Backpropagation technique.[6] In this study, we use the following equations to get the vector representations based on a stochastic gradient descent method:

$$\begin{aligned} \boldsymbol{v}_o{}^{new} = \boldsymbol{v}_o{}^{old} - \alpha \frac{\partial E}{\partial \boldsymbol{v}_o} &= \boldsymbol{v}_o{}^{old} - \alpha \frac{\partial E}{\partial \boldsymbol{u}_o} \frac{\partial \boldsymbol{u}_o}{\partial \boldsymbol{v}_o} \\ &= \boldsymbol{v}_o{}^{old} - \alpha \cdot \boldsymbol{e}_{out} \cdot \boldsymbol{h} \end{aligned} \tag{4}$$

$$\begin{aligned} \boldsymbol{v}_i{}^{new} = \boldsymbol{v}_i{}^{old} - \alpha \frac{\partial E}{\partial \boldsymbol{v}_i} &= \boldsymbol{v}_i{}^{old} - \alpha \frac{\partial E}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{v}_i} \\ &= \boldsymbol{v}_i{}^{old} - \alpha \sum_o \left(\frac{\partial E}{\partial \boldsymbol{u}_o} \frac{\partial \boldsymbol{u}_o}{\partial \boldsymbol{h}} \right) \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{v}_i} \\ &= \boldsymbol{v}_i{}^{old} - \alpha \sum_o \cdot \boldsymbol{e}_{out} \cdot \boldsymbol{v}_o \end{aligned} \tag{5}$$

where α is the learning rate which determines the amount of new information added into the estimation of the new weighting vector in this iteration; e_{Out} is the error vector that tracks the difference between output score and ground truth from last iteration; h is the intermediate vector in the hidden layer, where it is mapped from input vector, and then maps itself onto output vector $u_o = v_o^T h$.

Through these two stochastic gradient descent equations, the vector representation of an arbitrary input real estate description (v_i) can be obtained. Although the lengths of real estate descriptions may be different, their vector representations are the same size. Each v_i is a single–row vector, with a fixed number of columns (n–dimensional). By mapping house descriptions into an n-dimensional vector space (h), we can always obtain a fixed–length vector representation v_i.

The vector representation approach discussed in this section provides a technical ground for the creation of a numerical description uniqueness measure in the next subsection.

[6] A method to calculate a gradient based on the error estimation in the current iteration, which will be used for the calculation of the weighting vectors in Artificial Neural Networks.

2.2 Construction of the Uniqueness Measure

From the previous subsection, we obtained vector representations to quantify the information contents of house descriptions. We define the pairwise distance between two vectors to represent the relative semantic distance between the corresponding houses descriptions. This distance is measured using the angle between a pair of vectors obtained during the vectorization process (v_i in Equation 5), shown in the equation below.

$$Distance(\boldsymbol{v}_1, \boldsymbol{v}_2) = 1 - cos(\boldsymbol{v}_1, \boldsymbol{v}_2) = 1 - \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1|| \cdot ||\boldsymbol{v}_2||} \tag{6}$$

Notice that the distance defined in Equation 6 is the cosine distance between two vectors, where 0 means two identical descriptions with 0 semantic distance in between. This measure is mathematically bounded between 0 and 1.

Figure 3 provides a visual demonstration of the effectiveness of our ML algorithm. In the top text box, the query is "Lenox Mall", a shopping center in northern Atlanta. The blue pins on the map are houses related to the query. The middle text box displays the description of a selected house on the map. The bottom text box shows the most similar descriptions found in the data by the ML algorithm via Equation 6.[7] This figure shows our algorithm can successfully sort houses based on descriptions similarity/difference. In this particular example, all the similar houses are near the Lenox mall although the name of the mall does not directly show up in some of the descriptions.

Table 1 compares the pairwise semantic distances between the description of a subject houses with that of a few comparables in a neighborhood called the Grant Park subdivision. The distance 0 in the first row implies the description is being compared to itself. The pairwise distance between two descriptions increases as their semantic meanings deviate from each other. Notice that in the house descriptions, there are many abbreviations and typos. For instance, "granite" vs. "granit," "b'ful" vs. 'beautiful," "hrwds" vs. "hard–

[7]We conceal the house ID and the program copyright note to protect data privacy as well as to hide our names during the review process.

wood–floors," etc. The relationship between the paragraphs cannot be properly analyzed via a simple method based on keywords or word frequencies.

To assess the uniqueness of a description compared to its cohorts (within the same neighborhood in this paper), we compute the average pairwise distances from the house of interest to other houses, as shown in Equation 7 and Figure 5.

$$Unique_i = \frac{\sum_1^{N-1} (pairwise\, distance)}{(N-1)\ pair\, of\, houses} \tag{7}$$

Once the uniqueness scores have been obtained, we include this variable in our hedonic pricing model, which will be introduced in detail in Section 2.3.

2.3 Hedonic Pricing Model

We employ a classic linear hedonic model to estimate the impact of unique property descriptions on home prices. Our model controls for heterogeneous characteristics of houses, sale year, location, special transaction circumstances, as well as agent unobservables. The full empirical specification takes the following form:

$$Ln(Price_i) = \alpha + \theta Unique_i + X_i'\beta + \eta_z + \mu_c + \delta_t + \mu_c \times \delta_t + \varepsilon, \tag{8}$$

$Ln(Price_i)$ is the natural log of the sale price of house i. $Unique_i$ is the description uniqueness score derived from our machine learning semantic analysis model. X_i is a vector of physical characteristics and transaction circumstances of house i. The physical characteristics include number of bedrooms (Bed), square footage in hundred (Sqft), age (Age), number of fireplaces (Fireplaces), size of lot (Large Lot), whether the house has a pool (Pool), whether the house is recently renovated (Renovated), and whether the house comes with a special recreational feature such as access to a lake or a golf course (Feature). We also include dummy variables that indicate whether a sale has the following transaction circumstances: sold without a repair escrow (Sold–As–Is), sold by an agent who represents both the seller and the buyer

(Dual), and listing agent is the seller or is related to the seller (Owner Agent). η_z is a vector of listing agent fixed effects, δ_t is a vector of transaction year fixed effects, and μ_k is a vector of location fixed effects. Standard errors are clustered at the ZIP Code level.

The purpose of this empirical setup is to show that property descriptions reveal additional information that affect property sale prices besides the wide range of controls described in the paragraph above. $Unique_i$ is the variable of interest in our model. Our interest is in determining the sign and magnitude of the coefficient θ of $Unique_i$. A positive and statistical significant coefficient θ indicates unique property descriptions lead to price premiums, and a negative and statically significant coefficient θ indicates unique property descriptions lead to price discounts. Our null hypothesis is that the uniqueness of a real estate description has no effect on the sale price of a house: $H_0 : \theta{=}0$.

3 Data and Descriptive Statistics

The data we use encompass more than 40,000 single family home sales in Atlanta, GA from January 2010 to December 2017. The source of the data is the Multiple Listing Service (MLS). The information provided in the MLS data includes the address of each house, a wide range of house characteristics, critical dates regarding the transaction, unique IDs of the listing and buying agent, and most importantly the written description.

To identify neighboring houses, we geocoded the property addresses in our data and grouped houses based on their corresponding census blocks. Census blocks are small statistical areas bounded by visible features such as roads, streams, and railroad tracks, and by non–visible boundaries such as property lines, city, township, school district, county limits and short line-of-sight extensions of roads. In a city like Atlanta, a census block looks like a city block bounded on all sides by streets.[8] In Atlanta, houses in the same census block are most likely located in proximity and share the common infrastructures such as parks,

[8] "What are census blocks?"
https://www.census.gov/newsroom/blogs/random-samplings/2011/07/what-are-census-blocks.html

highways, and school districts. Figure 4 gives an overview of the geographical distribution of the data analyzed in this study.

We impose three restrictions on the descriptions data: First, we only include houses for which the property descriptions are longer than 9 characters. Second, we limit our sample to houses in census blocks with more than three sales during the sample period. Finally, we only include properties that were sold because descriptions of unsold houses are often deleted when a house was taken off the market.

The final data used in this study consists of 40,918 transactions: 37,124 unique sales and 3,794 repeat sales. We use the unique sales to deliver our baseline results and the repeat sales data for robustness tests. Table 2 displays a set of basic descriptive statistics for the data used in this study. The average home in our sample is 46 years old, has 2.7 bedrooms and 3.6 bathrooms. It is listed for $390,000 and is sold for $373,000 three and half months later. Since we only focus on the city, most of the houses sold in this area sit on small lots. Only 2.5 percent of the homes in our data are built on lots that are greater than one acre.

Table 3 displays a set of basic descriptive statistics for the description uniqueness score variable estimated by the Machine Learning algorithm. The average description in our data uses six sentences and 80 words to describe a house for sale. $Unique_i$ measures the semantic difference between the property description of house i and descriptions of neighboring houses sold during our sample period. This measure is bounded between 0 (a low level of semantic deviation) and 1 (a high level of semantic uniqueness). Neighboring houses are defined as homes sold within the same census blocks. $Unique_i$ clusters around 0.7 with the minimum value equals to 0.002 and its maximum value is 0.983.

Figure 4 shows the average uniqueness score by listing year, sale year, property age percentiles, and listing price percentiles. The average $Unique_i$ for the entire sample is 0.0726 and is shown in all four diagrams by the red dashed–line. The cohort average unique scores are displays with +/- one standard deviation. We do not observe any change in description uniqueness over the observation years, indicating there is no drastic change in description

language or home features over the observation period.

4 Empirical Results

In this section, we present empirical results reported by our hedonic models to explore the effects of $Unique_i$ on real estate sale prices. Table 4 displays a subset of coefficient estimates for our baseline specifications, in which we gradually add control variables. All of the specifications include Location FE, year FE, as well as Location×Time FE to eliminate spatial and temporal market impacts on house prices.

We begin our analysis by testing whether written real estate descriptions capture additional information than numerical house characteristics (results are shown in Column (1)). Controlling for a full set of physical covariates such as square footage, age, number of bedrooms, lot size, etc., we find description uniqueness is positively and statistically significantly correlated with sale prices, implying that houses advertised by unique property descriptions are sold for higher prices compared to houses sold in the same region of the city and under the same market condition. Since $Unique_i$ is bounded between 0 and 1, an economical meaningful interpretation of θ=0.597 is one standard deviation (0.08) increase in $Unique_i$ increases the sale price of a house by 4.8%. In addition, the coefficient estimates associated with the property characteristics are largely consistent with what previous studies have documented. For example, house age is negatively correlated with sale prices, whereas number of bedrooms, square footage, lot size, and the existence of a pool are positively related with sale prices.

Extant studies point out that certain transaction circumstances also affect property sale prices. In column (2) we extend our analysis and include indicator variables to control for transaction circumstances. Consistent with findings of previous studies, agent–owned houses are associated with higher sale prices than non-agent owned houses (Levitt and Syverson, 2008 and Rutherford and Yavas, 2005); and dual agency transactions are associated

with lower sale prices than sales in which different agents represent the seller and buyer (Han and Hong, 2016 and Brastow and Waller, 2013). In addition, houses without repair escrows are sold for lower prices than those with repair escrows. Controlling for special transaction circumstances, the $Unique_i$ coefficient estimate increases by 3.9 percentage points in magnitude and it is in the 95–percent confidence internal. A one standard deviation increase in $Unique_i$ (0.08) leads to a 5.09% increase in the sale price of a house. Thus, we reject the possibility that the price premium associated with unique property descriptions in Column (1) are caused by observed transaction circumstances.

Next, we demonstrate our ML approach complements extant studies that rely on frequently used keywords to capture house and transaction information that is not reported by numerical data but is present in descriptions. Following Levitt and Syverson (2008) and Rutherford and Yavas (2005), we include indicator variables of information revealing keywords in Column (3).[9] The coefficient estimate θ slightly dropped from 0.636 to 0.620 and is in the 99–percent confidence interval. The drop in the estimate magnitude is because words are components of a description, therefore our uniqueness measure already accounts for information conveyed by keywords. The small percentage change (2.5%) in the magnitude of θ highlights the difference of our approach compared to the widely used keyword–based approach: our uniqueness score gauge the semantic differences across different descriptions while keywords focus on common good/bad features mentioned in them.

The results so far suggest that sellers who want to achieve high sale prices should seek agents who provide unique property descriptions. One concern for the results documented in columns (1)–(3) in Table 4 is that $Unique_i$ might be spuriously correlated with unobserved heterogeneities in real estate agents. In Column (4), we include agent fixed effects to obtain a within agent estimator of $Unique_i$ on sale prices. The coefficient estimate of $Unique_i$ equals to 0.885 and is statistically significant suggesting that unique property descriptions are good marketing tools that work for all transactions. A one standard deviation increase in $Unique_i$

[9]We include the same words listed in Levitt and Syverson (2008) Table 1.

(0.08) leads to a 7% increase in the sale price of a house in our data, which is approximately $26,000. The empirical specification of Column (5) is similar to that in Column (4) except we drop the keyword indicators. The exclusion of keyword indicators lead to a small drop in the R^2 from 0.885 in Column (4) to 0.875 in Column (5), further demonstrating the information content captured by the unique score is beyond that conveyed by keywords. A one standard deviation increase in $Unique_i$ (0.08) leads to a 7.6% increase in the sale price of a house in our data, which is approximately $28,000.

Based on the results reported in Table 4, we reject the null hypothesis, H_0 : θ=0 in favor of the alternative hypothesis that $Unique_i$ captures the "soft information" numerical data unable to capture. In summary, threads of evidence presented in this section suggest unique real estate descriptions lead to higher sale prices, and this is not driven by agent unobservables.

4.1 Unique Feature Versus Unique Language

Like all advertisements, property descriptions disseminate a combination of facts and opinions. Therefore, the uniqueness of a property description might come from both the house itself as well as the marketing language used in its advertisement. While examples of unique features include special architectural designs or special recreational spaces, overly dramatic expressions would contribute to language uniqueness.[10]

To separate the price impact of uniqueness driven by agents marketing language from that driven by property features, we analyze how the change of description uniqueness between two sales ($\Delta Unique_i$) affect the prices of a house. The dependent variable in the repeat sales models is $\Delta Ln(Price_i)$ and we control for changes in physical and transaction covariates,

[10]For example, some agents create memorable names for houses such as "Biltmore Rose Cottage;" and some agents write dramatic sentences such as "You are humbled, like what happens when we stare out to sea and feel small" in property descriptions.
More examples can be found in "Put More Love Into Your Listing Ads" Daily Real Estate News, February 16th, 2016, `http://realtormag.realtor.org/daily-news/2016/02/16/put-more-love-your-listing-ads` and "Online Listing Descriptions Gone Wild" Daily Real Estate News, August 1st, 2013, `http://realtormag.realtor.org/daily-news/2013/08/01/online-listing-descriptions-gone-wild`

an indicator variable for renovation (Renovated), and changes in the House Price Index (ΔHPI) between two sales. Following Case and Shiller (1989), the coefficient estimate for $\Delta Unique_i$ measures the price impact of language uniqueness in the repeat sales model. Recall the coefficient estimate for $Unique_i$ captures the combined price impact of both language uniqueness and feature uniqueness in the baseline model.

The results of our repeat sales analysis are presented in Column (1)–Column (4) in Table 5. Column (1)–(2) control for county fixed effects and Column (3) and Column (4) control for MLS area fixed effects. Column (1) and (3) control for listing agent fixed effects and Column (2) and (4) control for listing office fixed effects. The coefficient estimate β is only weakly significant in the second repeat sales model with a low R^2 (0.472).

Columns (5)–(8) display the full sample analysis results that correspond to empirical specifications in Columns (1)–(4) to compare the magnitude difference between price impacts of language uniqueness and overall (language and feature) uniqueness. The regression coefficient for $Unique_i$ in Column (6) equals to 1.366 and is 5.8 times larger than the estimate of $\Delta Ln(Price_i)$ in Column (2), implying language uniqueness may account for **at most** 17% of the 7.8% price premium associated with the overall description uniqueness.

In summary, the evidence from this section suggests unique house features introduced in the descriptions are key determinants of higher sale prices documented by our principle findings, especially given the low R^2 and low statistical significance of θ in Column (2) in Table 5.

4.2 Good Uniqueness Versus Bad Uniqueness

The uniqueness measure $Unique_i$ captures the semantic deviation of a property description from its neighborhood average. Because likable and unlikable features can both increase the uniqueness of a house, in this section, we test whether good uniqueness and bad uniqueness impose the same level of effect on sale prices. We create indicator variables for the superior and inferior houses by whether positive or negative keywords are used in their descriptions.

Examples of the positive words are "landscaped," "move–in," "brand new," etc. Examples of the negative words are "needs updating," "foreclosure," "TLC," etc.

Table 6 show empirical estimations of a difference–in–difference model. We identify a house to be a bad one if at least one negative word appears in its description. Due to the high frequency of good keywords used in the descriptions, we restrict good houses to those that have more than five positive keywords in their descriptions. This strategy identified 5,372 bad houses and 7,546 good ones in our sample, which account for the bottom 13% and top 19% of the sample respectively. We also excluded 699 observations that meet the standards for both good and bad houses from the testing sample to avoid biases introduced by word misuses.

Column (2) of Table 6 provides empirical evidence that good (superior) houses are associated with above average sale prices. The coefficient estimate for Good$\times Unique_i$ is -0.66 and lies in the 90–percent confidence interval, implying the superior (good unique) houses suffer from price discounts when the neighboring houses are lower in quality. This effect can be interpreted as negative spillovers of the less–attractive houses in proximity.

Column (2) also shows bad (inferior) houses are sold for below market average prices. The coefficient estimate for the interaction term Bad$\times Unique_i$ is 1.5 and lies in the 99–percent confidence interval, implying bad houses in good neighborhoods are associated with price premiums due to the positive spillovers from the more–attractive surrounding houses.

Taken together, although the inferior houses were sold for lower prices than the superior houses, being inferior compared to the neighbors mitigates negative price impacts caused by undesirable house characteristics.

4.3 Unique Description and Real Estate Liquidity

In this section, we study the impact of uniqueness on the liquidity of houses. The dependent variable of all the empirical specifications in this section is the number of days a property stays on the market before being sold (DOM).

Columns (1)–(2) in Table 7 employ the same control variables as Columns (4)–(5) in Table 4 including numerical home features, transaction circumstances, time and location fixed effects, and agent fixed effects. The coefficient estimate for $Unique_i$ is 28.8 and lies in the 99–percent confidence interval, implying a one standard deviation increase in $Unique_i$ (0.08) postpones the sale time of a house by approximately 2.3 days.

Extant studies show real estate prices are positively correlated with days–on–market (Hendel et al., 2009, Levitt and Syverson, 2008, Rutherford and Yavas, 2005). Following Levitt and Syverson (2008), Columns (3)–(4) simultaneously model DOM and real estate listing prices. The magnitude and statistical significance of the $Unique_i$ estimate reported by the joint models are comparable with those reported in Columns (1)–(2).

Taken together, the results shown in this section suggest unique houses take longer to sell, which is consistent with the findings of Haurin (1988). A one standard deviation increase in uniqueness leads to approximately 2.3 days delay in the sale time.

5 Unique Description and Realtor Experiences

Evidence shown in the previous section suggests our ML analysis of real estate descriptions bridges the information gap between structured real estate data and the houses by dissipating information about unique house features. In this section, we provide a simple application of our uniqueness measure on the analysis of whether market experiences improve real estate agents' ability to write effective advertisements.

Previous studies suggest experiences may improve real estate agents' sales skills (Han and Strange, 2015, Levitt and Syverson, 2008, Rutherford and Yavas, 2005). Because the MLS systems impose a maximum length on the descriptions, the listing agents must use their market knowledge to select house features that signal the most attractive information to the buyers. Therefore, descriptions written by agents who are more experienced might be more effective than those written by inexperienced agents.

Figure 7 shows the relationship between description uniqueness ($Unique_i$) and the listing agents' market experiences. The y–axis plots the uniqueness of each real estate description, and the x–axis shows the total number of houses the corresponding listing agent has sold before writing the particular advertisement. The solid horizontal line implies there is no statistical trend between the uniqueness of a description and the listing agent's market experiences.

We then test for potential heterogeneous impacts of unique descriptions written by agents with different levels of experiences, estimating the following linear model:

$$Outcome = \alpha + \theta_1 Unique_i + \theta_2 TotalSale + \theta_3 TotalSale \times Unique_i + X_i'\beta + \eta_z + \mu_c + \delta_t + \mu_c \times \delta_t + \varepsilon \tag{9}$$

where $Unique_i$ is the description uniqueness score, $TotalSale$ is the number of houses an agent has sold previously when selling the current house, and $TotalSale \times Unique_i$ is interaction of the first two variables. We also control for physical characteristics, transaction circumstances, location and time fixed effects, and most importantly, agent fixed effects.

The dependent variable in Table 8 Columns (1)–(2) is $Ln(Price_i)$. The coefficient estimates of $Unique_i$ remain positive and statistically significant. The negative and statistical significant estimates associated with $TotalSale$ suggest the sale price of a house decreases by 0.2% with one more count of total homes sold by the listing agent. This finding is consistent with results of Bian et al. (2015), suggesting the amount of effort devoted to the selling of each house decreases as the selling agents' experiences increase. The estimated effect of $TotalSale \times Unique_i$ in Columns (1)–(2) is positive and statistically significant, implying agents learn from past transactions and become more experienced in selecting value enhancing house features to advertise in descriptions. With one more prior home sale experience, description uniqueness increases the sale price of a house by 0.21%.[11]

Turning to the liquidity analysis results presented in Columns (3)–(4) in Table 8, the

[11] $\theta_3 \times$ Average $(Unique_i) = 0.003 \times 0.7 = 0.21\%$

coefficient estimates for the interaction term $TotalSale \times Unique_i$ are statistically insignificant. This finding is especially striking given the estimated effect of the interaction on sale prices is positive. Recall we have shown in Section 4.3 that houses advertised with unique descriptions must stay on the market longer to receive higher sale prices, the statistically insignificant impact of $TotalSale \times Unique_i$ on DOM imply market experiences help agents to effectively advertise houses by their value enhancing unique characteristics without delaying the closing dates.

In summary, the results in this section show sales experiences have a positive impact on agents' skills. Although experienced agents devote less effort to sell houses than inexperienced agents, their market knowledge derived from experiences enables them to write property descriptions more effectively.

6 Conclusion

Extant literature has documented "soft" information plays a crucial role in the price determination process of highly differentiable goods, such as houses. In this study, we investigate the price impact uniqueness, a type of "soft" information captured in real estate advertisements.

Our contribution is threefold. First, we proposed a Machine Learning algorithm to quantify the semantic uniqueness of textual property descriptions; Second, we estimated the impact of description uniqueness on real estate sale prices and marketing time using linear hedonic pricing models. A one standard deviation increase in description uniqueness is associated with a 5.6% increase in sale prices while delaying the closing time by 2.3 days. A large fraction of the price premium associated with description uniqueness is caused by unique features of the underlying houses while the impact of language uniqueness is limited. Finally, we offer a discussion on the relationship between description uniqueness and market experiences. Our findings suggest agents' ability to write effective property descriptions increases in his or her previous sales experiences.

This study also provides several theoretical and empirical insights in response to the emerging Artificial Intelligence technologies for economics research. First, our ML algorithm defines the meanings of words within their contexts, overcoming a common limitation of the keyword–based textual analysis methods: "simply counting words (bag–of–words) ignores important context and background knowledge (Larcker and Zakolyukina, 2012)." Our Machine Learning algorithm naturally preserves the meaning of words within their contexts, and therefore, can understand the nuances of marketing language as well as unstandardized abbreviations in property descriptions. Second, the ML–Hedonic approach used in this study provides an example of the integration of unsupervised learning methods into economic analysis.

A recent boom in Artificial Intelligence and Machine Learning has made drastic impacts on academic research. While most of the recent ML studies in economics and finance focus on predictions, this paper suggests ML has also allowed the hardening of "soft" information using textual data. The ML–Hedonic approach in this paper allowed us to draw economic inferences about the impact of real estate description uniqueness on sale prices. In addition, the context–based ML algorithm can be applied to extract information from a wide range of textual documents to study a variety of economic phenomena.

References

Bernstein, A., M. Gustafson, and R. Lewis (2018). Disaster on the horizon: The price effect of sea level rise. *Journal of Financial Economics*.

Bian, X., B. D. Waller, G. K. Turnbull, and S. A. Wentland (2015). How many listings are too many? agent inventory externalities and the residential housing market. *Journal of Housing Economics 28*, 130 – 143.

Brastow, R. and B. Waller (2013). Dual agency representation: Incentive conflicts or efficiencies? *Journal of Real Estate Research 35*(2), 199–222.

Case, K. E. and R. J. Shiller (1989). The efficiency of the market for single-family homes. *American Economic Review 79*(1), 125–137.

Dai, A. M., C. Olah, and Q. V. Le (2015). Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance 68*(3), 1267–1300.

Garmaise, M. J. and T. J. Moskowitz (2004). Confronting information asymmetries: Evidence from real estate markets. *The Review of Financial Studies 17*(2), 405437.

Goodwin, K., W. B. W. H. S. (2014). The impact of broker vernacular in residential real estate. *Journal of Housing Research 23*(2), 143–161.

Han, L. and S.-H. Hong (2016). Understanding in-house transactions in the real estate brokerage industry. *The RAND Journal of Economics 47*(4), 1057–1086.

Han, L. and W. C. Strange (2015). Chapter 13 - the microstructure of housing markets: Search, bargaining, and brokerage. *5*, 813 – 886.

Haurin, D. (1988). The duration of marketing time of residential housing. *Real Estate Economics 16*(4), 396–410.

Hendel, I., A. Nevo, and F. Ortalo-Magn (2009, December). The relative performance of real estate marketing platforms: Mls versus fsbomadison.com. *American Economic Review 99*(5), 1878–98.

Larcker, D. F. and A. A. Zakolyukina (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research 50*(2).

Levitt, S. D. and C. Syverson (2008). Market distortions when agents are better informed: The value of information in real estate transactions. *The Review of Economics and Statistics 90*(4), 599–611.

Liberti, J. M. and M. A. Petersen (2018). Information: Hard and soft. *Working Paper*.

Lindenthal, T. (2017). Beauty in the eye of the home-owner: Aesthetic zoning and residential property values. *Real Estate Economics*.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance 66*(1), 35–65.

Muehlenbachs, L., E. Spiller, and C. Timmins (2015). The housing market impacts of shale gas development. *The American Economic Review 105*(12), 3633–3659.

Mullainathan, S. and J. Spiess (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives 31*(2).

Nowak, A. and P. Smith (2017, June). Textual Analysis in Real Estate. *Journal of Applied Econometrics 32*(4), 896–918.

Palmquist, R. B. (1984). Estimating the demand for the characteristics of housing. *The Review of Economics and Statistics 66*(3), 394–404.

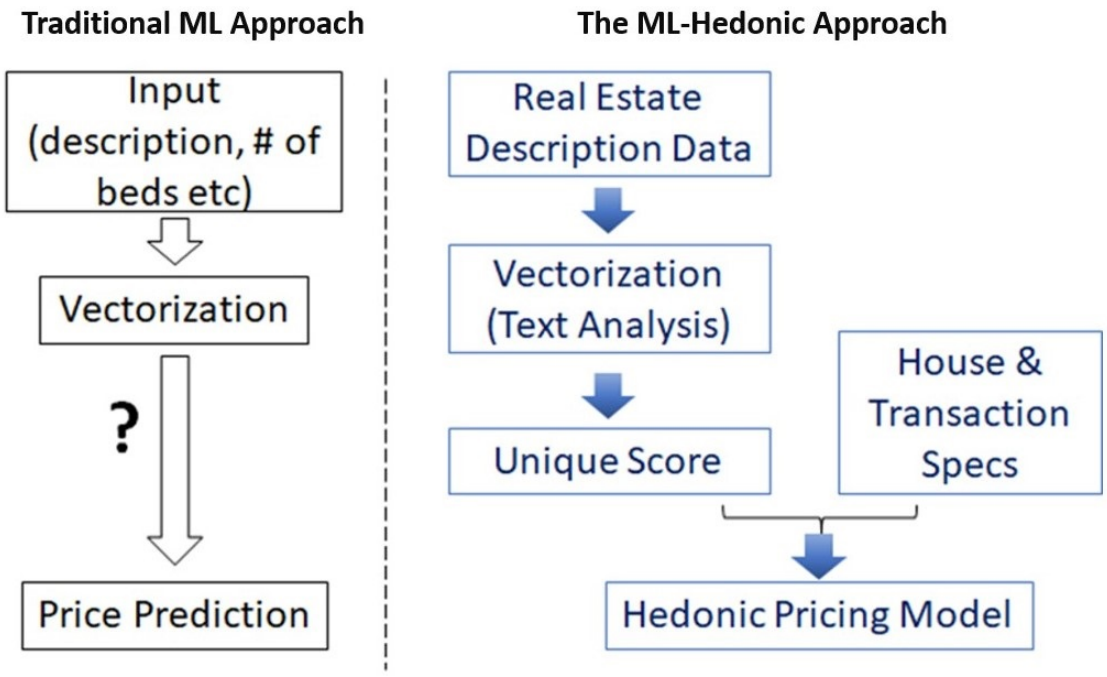
Pryce, G., . O. S. (2008). Rhetoric in the language of real estate marketing. *Housing Studies 23*(2), 319–348.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy 82*(1), 34–55.

Rutherford, R. C., T. M. S. and A. Yavas (2005). Conflicts between principals and agents: evidence from residential brokerage. *Journal of Financial Economics 76*(3), 627 – 665.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance 62*(3), 1139–1168.

Figure 1: Traditional Machine Learning Approach versus the ML–Hedonic Approach



Notes: This figure summarizes the fundamental differences between real estate price analysis using supervised machine learning algorithms and our machine learning–hedonic hybrid approach used in this paper. Supervised ML algorithms predict house prices using both numeric and text data. In our approach, un–supervised machine learning algorithm is used to only analyze text data to calculate a uniqueness measure of each house($Unique_i$). The unique score then becomes a covariate in our hedonic pricing model.

Figure 2: Schematic Algorithm

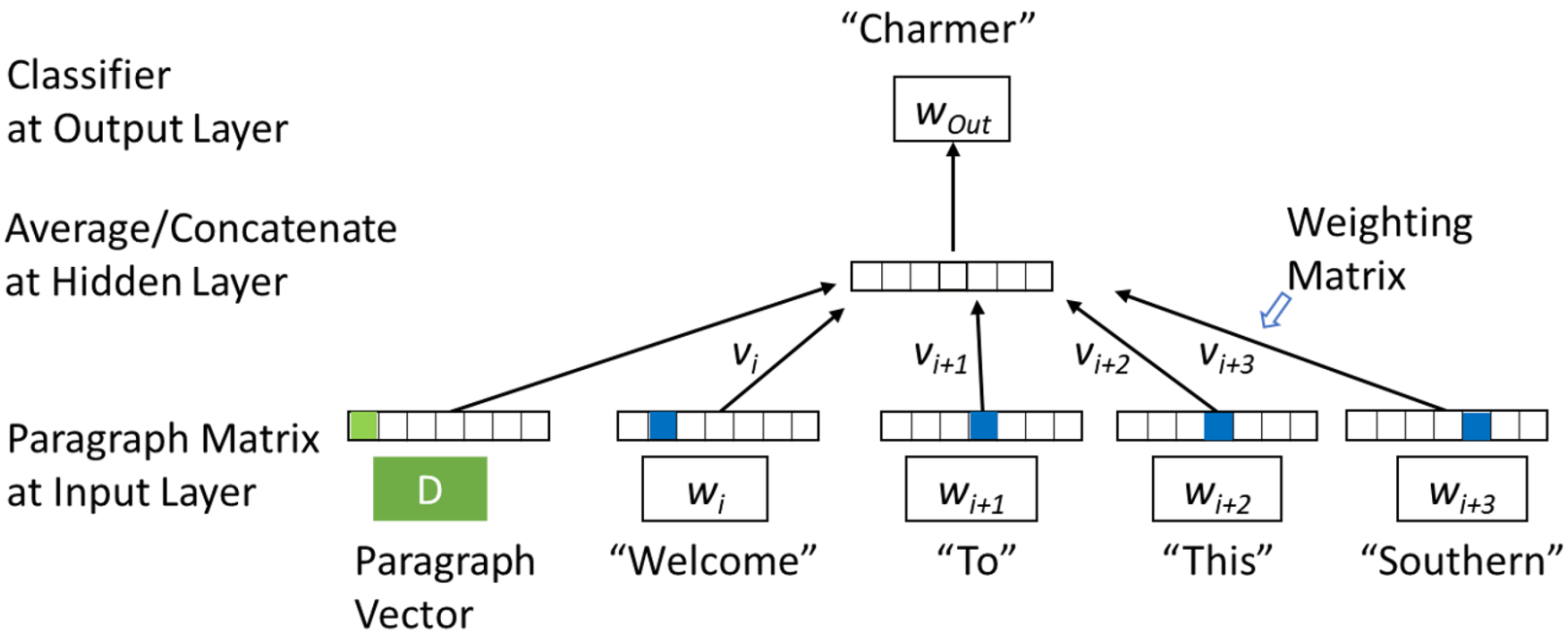
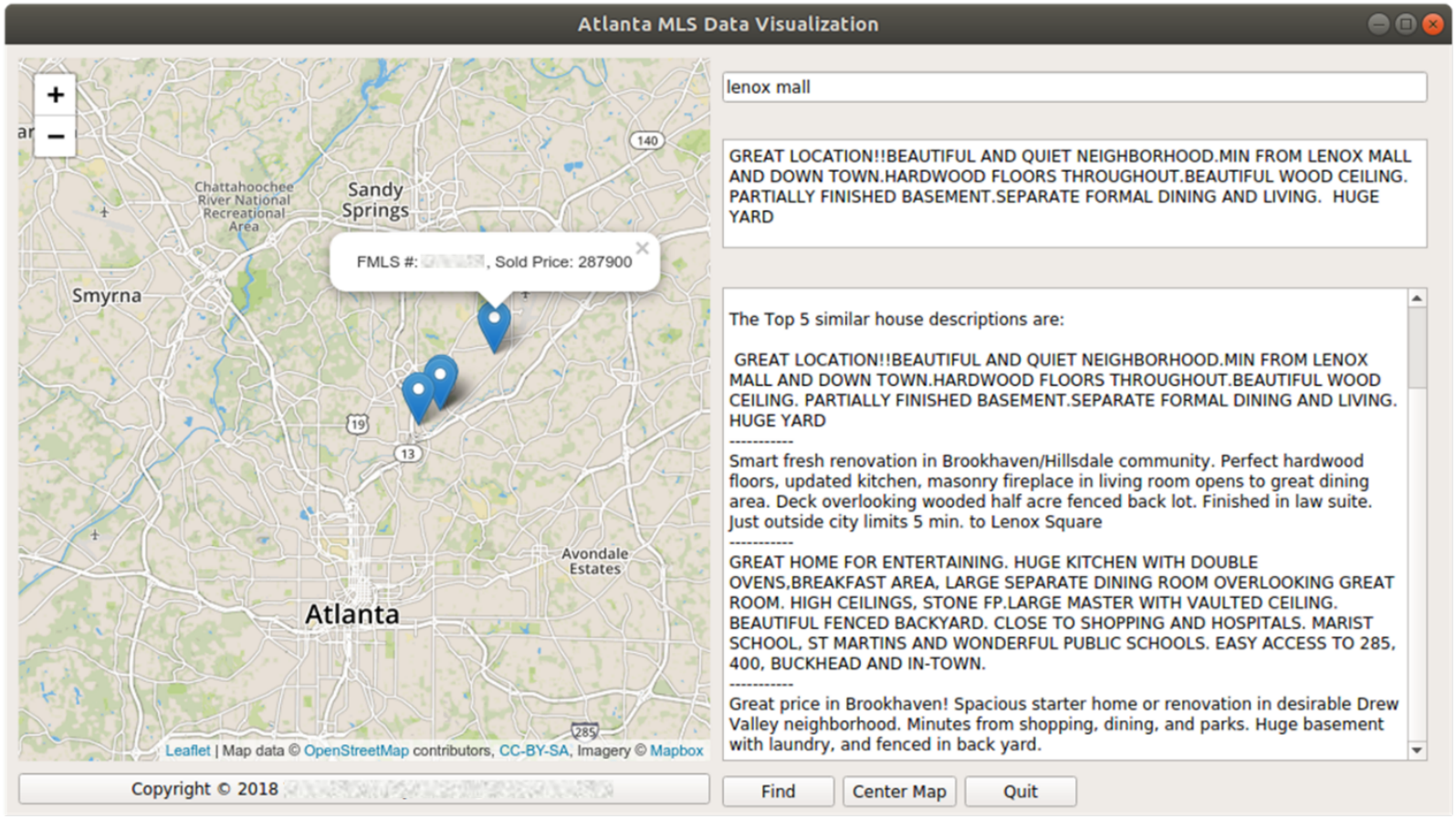
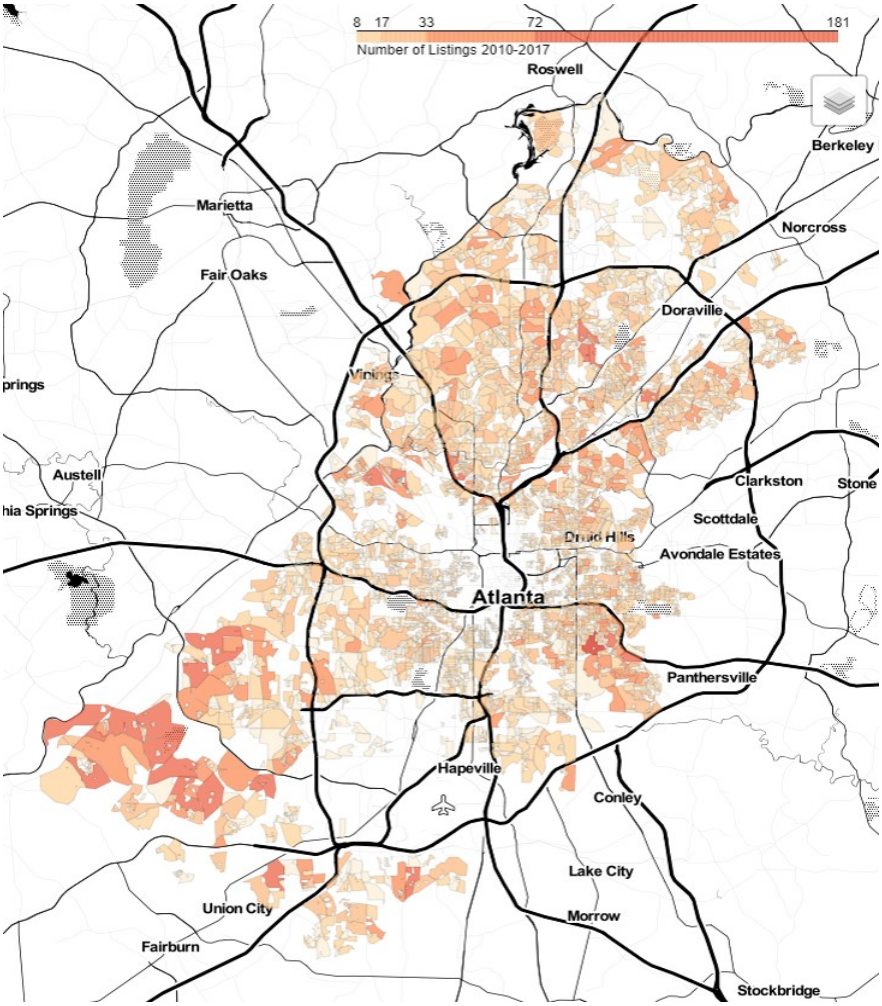


Figure 3: Algorithm Visualization



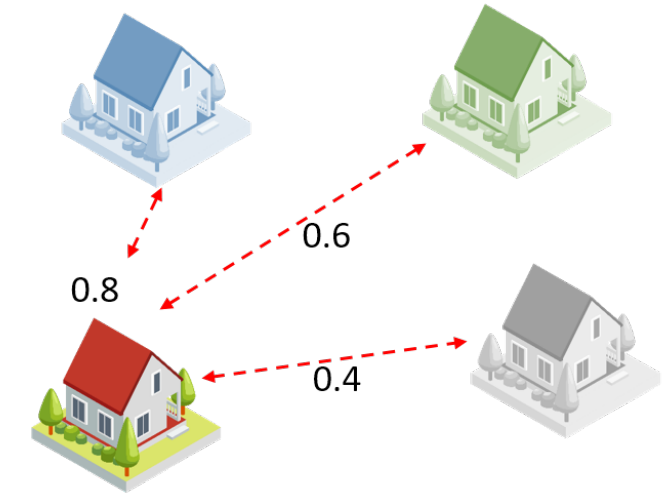
Notes: This figure provides a visual demonstration of our ML algorithm. In the top text box, the query is "Lenox Mall", a shopping center in northern Atlanta. The blue pins on the map are houses that are related to the query. The middle text box displays the description of a selected house on the map. The bottom text box shows the most similar descriptions found in the data by the ML algorithm. We conceal the house ID and the program copyright note for data privacy as well as to hide the our names during the review process. In this particular example, all the selected houses are near the Lenox mall although the name of the mall does not directly show up in some of the descriptions.

Figure 4: Geographical Distribution of the Real Estate Sales Sample



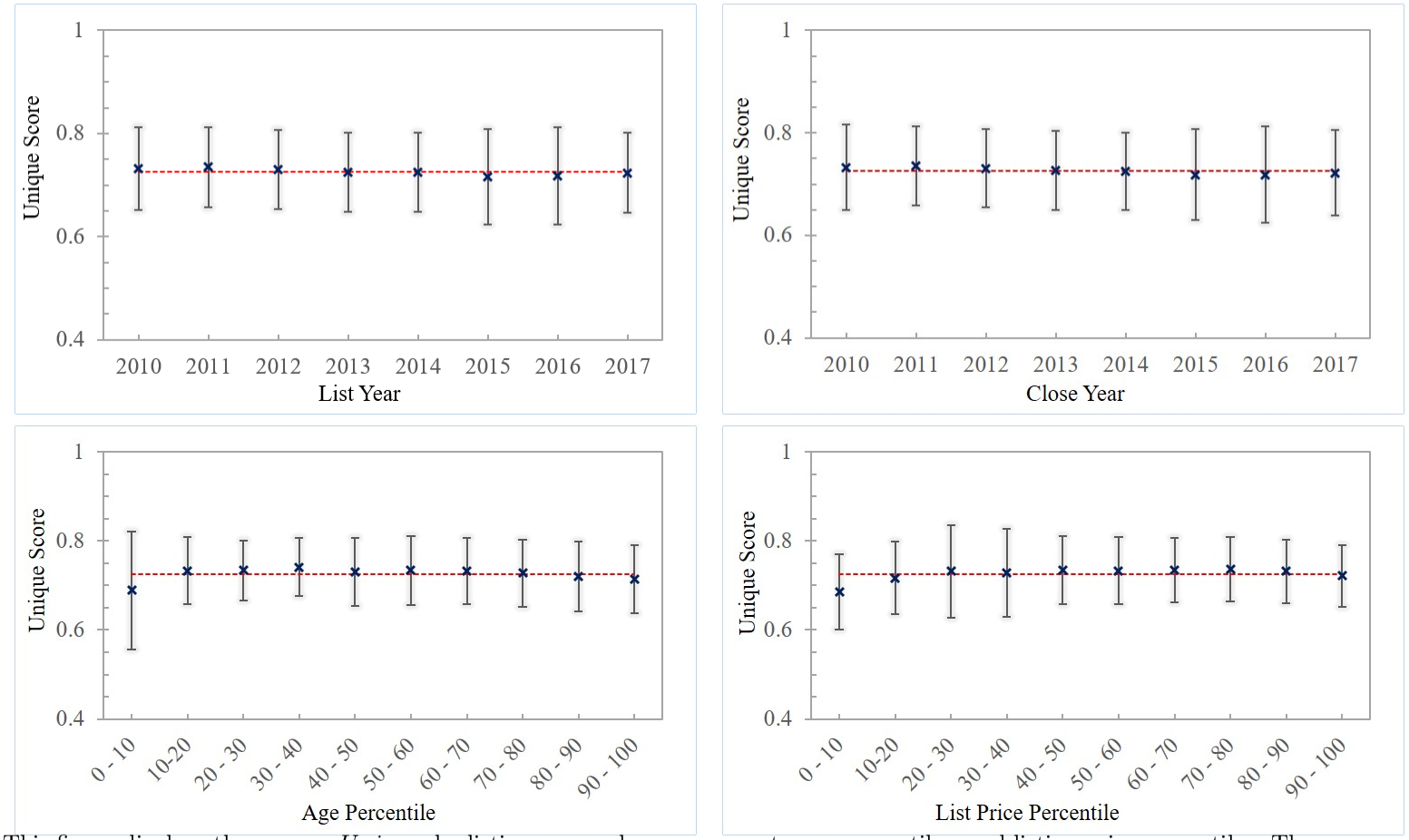
Notes: This figure displays an overview of the geographical distribution of the 40,918 single family houses analyzed in this study. Since we only focus on the city of Atlanta, most of houses sold in this area sit on small lots.

Figure 5: Schematic Unique Score Computation within a Neighborhood



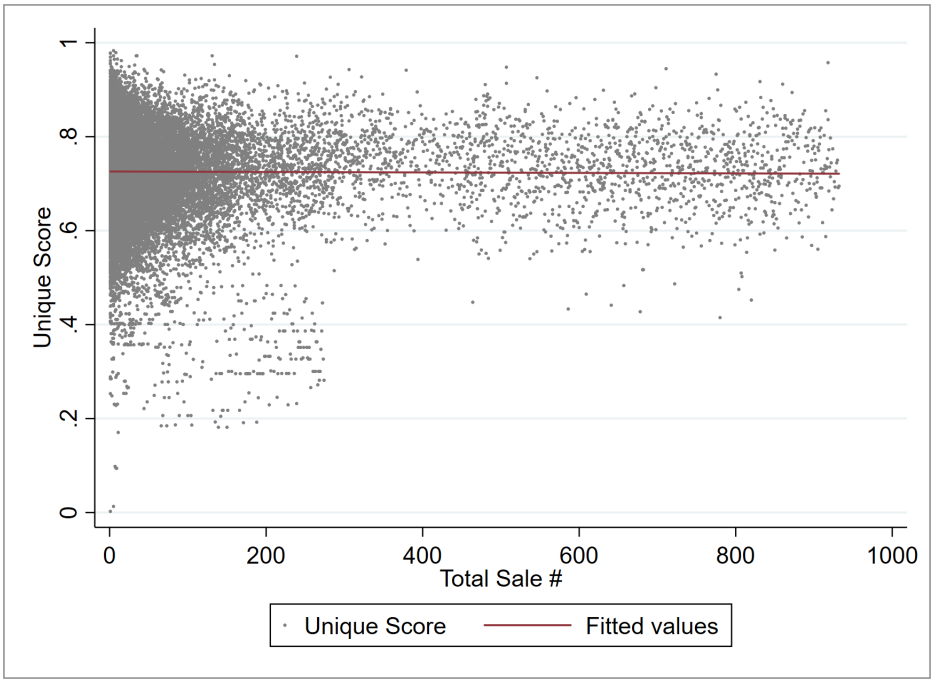
Notes: This figure displays an schematic computation of unique score for a house compared to its cohorts within the same neighborhood. All numbers are provided for illustrative purpose.

Figure 6: Average Unique Score by Different Cohorts versus Sample Average



Notes: This figure displays the average $Unique_i$ by listing year, sale year, property age percentiles, and listing price percentiles. The average $Unique_i$ for the entire sample is 0.0726 and is shown in all four diagrams by the red dashed–line. The cohort average unique scores are displays with +/- one standard deviation.

Figure 7: Unique Score vs. Agent experience (Total Number of Sales)



Notes: This figure shows the relationship between description uniqueness ($Unique_i$) and the listing agents' market experience. The y–axis plots the uniqueness of each real estate description, and the x–axis shows the total number of houses the corresponding listing agent has sold before writing this advertisement. The solid horizontal line implies there is no statistical trend between the uniqueness of descriptions and listing agents' market experience.

Table 1: Property Descriptions by Pairwise Distances

Subject Description	Comparable Description	Distance(Subj. vs. Comp.)
all new systems–easy loan qual, no repairs necessary! motivated seller–bring all offers!! welcoming 2 story covered dual porch. shows like a model! b'ful hrdwds extra trim on main. Elegant din rm. french doors to office bonus rm! open kit w granite stainless! lrg mstr w frpl priv deck! fireside liv rm!	all new systems–easy loan qual, no repairs necessary! motivated seller–bring all offers!! welcoming 2 story covered dual porch. shows like a model! b'ful hrdwds extra trim on main. Elegant din rm. french doors to office bonus rm! open kit w granite stainless! lrg mstr w frpl priv deck! fireside liv rm!	0
all new systems–easy loan qual, no repairs necessary! motivated seller–bring all offers!! welcoming 2 story covered dual porch. shows like a model! b'ful hrdwds extra trim on main. Elegant din rm. french doors to office bonus rm! open kit w granite stainless! lrg mstr w frpl priv deck! fireside liv rm!	newer construction craftsman style charmer in awesome grant park location! absolutely adorable curb appeal! *attractive dbl frt porch beautiful hdwds desirable flr plan huge, light filled fam rm w cozy frpl elegant din rm w detailed moldings gorgeous kit w island bfast rm convenient powder rm on main luxurious mstr ste w frpl balcony access incredible 3rd lvl ideal for rec rm or teen ste det 2–car gar walk to grant pk, zoo atl, turner field stanton elem! purchase for as little as 5% down–apprvd for homepath mortgage renovation financing!	0.412
all new systems–easy loan qual, no repairs necessary! motivated seller–bring all offers!! welcoming 2 story covered dual porch. shows like a model! b'ful hrdwds extra trim on main. Elegant din rm. french doors to office bonus rm! open kit w granite stainless! lrg mstr w frpl priv deck! fireside liv rm!	grant park's in-town living! this beautiful 2story hm has a back deck off great rm, sep dining, hrdwds, mud rm, lots of windows, granit kit, oversized cabinets, fp wood or gas, great yard, master w huge.	0.597
all new systems–easy loan qual, no repairs necessary! motivated seller–bring all offers!! welcoming 2 story covered dual porch. shows like a model! b'ful hrdwds extra trim on main. Elegant din rm. french doors to office bonus rm! open kit w granite stainless! lrg mstr w frpl priv deck! fireside liv rm!	foreclosure – two story brick front on bsmnt, prvt fenced yard, hrwds, sep liv din rm, frplc in fam rm and eat-in kitchen. easy showings....	0.639

Notes: This table compares the pairwise semantic distances between the description of a subject houses with that of a few comparables in a neighborhood called the Grant Park subdivision. The distance 0 in the first row implies the subject description is being compared to itself. The pairwise distance between two descriptions increases as their semantic meanings deviate from each other. Notice there are many abbreviations and typos in the descriptions. For instance, "granite" vs. "granit," "bful" vs. "beautiful," "hrwds vs. "hard–wood-floors," etc. The relationship between the paragraphs cannot be properly analyzed via a simple method based on keywords or word frequencies.

Table 2: Descriptive Statistics: Real Estate Sale Sample

VARIABLES	(1) Mean	(2) SD
DOM (number of days on market)	104.6	94.12
Listing Price ($ thousand)	390.0	451.9
Ln (Sale Price)	12.26	1.234
Sale Price ($ thousand)	373.2	411.6
Age	46.20	31.02
Fireplace (number of fireplace)	1.043	1.039
Sqft (hundred)	22.93	13.57
Bath (number of bathroom)	2.700	1.285
Bed (number of bedroom)	3.606	1.034
Listing Year	2,013	2.289
Sold Year	2,014	2.208
Ranch (indicator variable if ranch style)	0.440	0.496
Pool (indicator variable)	0.0491	0.216
Renovated (indicator variable)	0.0708	0.257
Sold-As-Is (indicator variable)	0.118	0.323
Auction (indicator variable if foreclosure auction)	0.0249	0.156
Large Lot (indicator variable if lot>=1 acre)	0.0247	0.155
Feature (indicator variable)	0.0134	0.115
Owner Agent (indicator variable if agent related to owner)	0.0285	0.166
Dual (indicator variable if dual agent)	0.0571	0.232

Note: This sample contains 40,918 single-family home sales in Atlanta from January 2010 to December 2017. The final data used in this study include 37,124 unique sales and 3,794 repeat sales. We use the unique sales to deliver the baseline results and the repeat sales data are used to conduct robustness analysis.

Table 3: Descriptive Statistics: Real Estate Description Sample

VARIABLES	(1) Mean	(2) Min	(3) Max	(4) SD
Word (# of words in property description)	80.27	10	164	31.50
Sentence (# of sentences in description)	5.900	1	23	2.778
Sell_Num (# of sales in census block)	22.60	3	181	26.09
$Unique_i$ (unique score)	0.726	0.00244	0.983	0.0827

Note: This sample contains 40,918 single-family home sales in Atlanta from January 2010 to December 2017. The final data used in this study include 37,124 unique sales and 3,794 repeat sales. We use the unique sales to deliver the baseline results and the repeat sales data are used to conduct robustness analysis.

Table 4: Uniqueness and Real Estate Sale Prices

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5
$Unique_i$	0.597*	0.636**	0.620***	0.885***	0.946***
	(0.298)	(0.272)	(0.206)	(0.185)	(0.217)
Age	-0.014***	-0.013***	-0.011***	-0.010***	-0.011***
	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Age×Age	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Bed	-0.041	-0.036	-0.017	0.002	-0.006
	(0.036)	(0.033)	(0.027)	(0.018)	(0.021)
Bath	0.240***	0.232***	0.216***	0.178***	0.190***
	(0.027)	(0.024)	(0.018)	(0.013)	(0.016)
Ranch	-0.268***	-0.251***	-0.213***	-0.157***	-0.175***
	(0.053)	(0.049)	(0.040)	(0.027)	(0.029)
Renovated	0.249***	0.208***	0.081***	0.099***	0.166***
	(0.047)	(0.043)	(0.024)	(0.015)	(0.025)
Sqft	0.009***	0.009***	0.009***	0.008***	0.009***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Large Lot	0.019	0.035	0.067**	0.086***	0.074**
	(0.046)	(0.045)	(0.033)	(0.025)	(0.030)
Pool	0.060**	0.067***	0.102***	0.114***	0.098***
	(0.025)	(0.024)	(0.017)	(0.016)	(0.019)
Feature	0.134**	0.153**	0.172***	0.127***	0.112**
	(0.061)	(0.060)	(0.060)	(0.046)	(0.045)
Sold–As–Is		-0.427***	-0.281***	-0.124***	-0.200***
		(0.063)	(0.041)	(0.024)	(0.030)
Owner Agent		0.089***	0.054***	0.039*	0.039
		(0.021)	(0.016)	(0.022)	(0.025)
Dual		-0.215***	-0.165***	-0.142***	-0.168***
		(0.031)	(0.024)	(0.020)	(0.023)
Constant	11.538***	11.535***	11.378***	11.391***	11.312***
	(0.226)	(0.210)	(0.192)	(0.210)	(0.223)
Observations	37,124	37,124	37,124	37,124	37,124
R-squared	0.764	0.777	0.806	0.885	0.875
House Characteristics	Yes	Yes	Yes	Yes	Yes
Keywords	No	No	Yes	Yes	No
Transaction Characteristics	No	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Location FE	Yes	Yes	Yes	Yes	Yes
Location×Year FE	Yes	Yes	Yes	Yes	Yes
Agent FE	No	No	No	Yes	Yes

Note: The dependent variable in all specifications is Ln(sale price). Robust standard errors are clustered at the ZIP Code level, shown in parentheses (*** p<0.01, ** p<0.05, * p<0.1)

Table 5: Unique Characteristics Versus Unique Language

VARIABLES	Repeat Sale				Full Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\Delta Unique_i$	0.165	0.236*	0.166	0.199				
	(0.174)	(0.134)	(0.170)	(0.131)				
$Unique_i$					1.377***	1.366***	1.001***	0.938***
					(0.205)	(0.247)	(0.222)	(0.239)
Age					-0.007**	-0.003	-0.011***	-0.011***
					(0.003)	(0.003)	(0.002)	(0.002)
Renovated	0.204***	0.237***	0.181***	0.215***	0.133***	0.120***	0.164***	0.187***
	(0.042)	(0.028)	(0.039)	(0.023)	(0.026)	(0.038)	(0.025)	(0.039)
Sqft					0.010***	0.012***	0.008***	0.008***
					(0.002)	(0.002)	(0.002)	(0.002)
Sold–As–Is					-0.216***	-0.463***	-0.197***	-0.360***
					(0.032)	(0.046)	(0.035)	(0.051)
ΔAge	-0.002**	-0.003***	-0.002**	-0.003***				
	(0.001)	(0.001)	(0.001)	(0.001)				
ΔHPI	0.009***	0.007***	0.010***	0.008***				
	(0.002)	(0.002)	(0.002)	(0.002)				
Constant	0.869***	0.903***	0.562*	0.246	10.326***	10.229***	10.919***	11.018***
	(0.256)	(0.154)	(0.292)	(0.149)	(0.165)	(0.212)	(0.242)	(0.252)
Observations	3,785	3,785	3,785	3,785	40,909	40,909	40,909	40,909
R-squared	0.746	0.472	0.761	0.538	0.827	0.717	0.868	0.807
House Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Transaction Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location FE	County	County	MLS Area	MLS Area	County	County	MLS Area	MLS Area
Location×Year FE	No	No	No	No	No	No	No	No
Agent FE	Yes	No	Yes	No	Yes	No	Yes	No
Listing Office FE	No	Yes	No	Yes	No	Yes	No	Yes
Cluster	ZIP Code	ZIP Code	ZIP Code	ZIP Code	ZIP Code	ZIP Code	ZIP Code	ZIP Code

Notes: The average change in uniqueness $\Delta Unique_i$ is -0.0007159 with a minimum of -0.29189 and maximum is 0.3818916. The standard deviation is 0.0749422.

Table 6: Good Uniqueness Versus Bad Uniqueness

VARIABLES	(1) Model 1	(2) Model 2
$Unique_i$	0.836***	0.900***
	(0.199)	(0.229)
Good (# of Positive word>=5)	0.317*	0.563**
	(0.176)	(0.225)
Good×$Unique_i$	-0.659**	-0.656**
	(0.249)	(0.297)
Bad (# of Negative Word>=1)	-1.297***	-1.561***
	(0.241)	(0.288)
Bad×$Unique_i$	1.576***	1.550***
	(0.310)	(0.355)
Constant	11.549***	11.545***
	(0.174)	(0.189)
Observations	32,634	32,634
R-squared	0.890	0.884
House Characteristics	Yes	Yes
Keywords	Yes	No
Transaction Characteristics	Yes	Yes
Year FE	Yes	Yes
Location FE	Yes	Yes
Agent FE	Yes	Yes
Cluster	ZIP Code	ZIP Code

Notes: This table displays the estimation results of

$$Ln(Price_i) = \alpha + \theta_1 Unique_i + \theta_2 Good + \theta_3 Good \times Unique_i + \theta_4 Bad + \theta_5 Bad \times Unique_i + X_i'\beta + \eta_z + \mu_c + \delta_t + \mu_c \times \delta_t + \varepsilon.$$

Standard errors are clustered at the ZIP Code Level (*** p<0.01, ** p<0.05, * p<0.1).

Table 7: Description Uniqueness and Real Estate Liquidity

VARIABLES	(1) DOM	(2) DOM	(3) DOM+Listing$	(4) DOM+Listing$
$Unique_i$	28.023***	28.789***	29.219***	28.854***
	(7.545)	(8.177)	(7.463)	(8.014)
Age	-0.199**	-0.188*	-0.214**	-0.189*
	(0.098)	(0.107)	(0.096)	(0.103)
Bath	3.933**	4.217***	4.197***	4.231***
	(1.496)	(1.525)	(1.421)	(1.462)
Ranch	-2.501**	-1.837	-2.721**	-1.850
	(1.131)	(1.167)	(1.216)	(1.286)
Renovated	-4.368**	-4.709**	-4.236**	-4.698**
	(1.892)	(1.969)	(1.903)	(2.005)
Sqft	0.361**	0.389**	0.374**	0.389**
	(0.173)	(0.166)	(0.173)	(0.165)
Large Lot	16.900***	18.415***	17.044***	18.421***
	(3.543)	(3.671)	(3.540)	(3.660)
Owner Agent	-8.522**	-7.594*	-8.471**	-7.591*
	(3.987)	(4.027)	(3.978)	(4.015)
Dual	7.566**	7.888**	7.401**	7.878**
	(3.132)	(3.198)	(3.106)	(3.171)
Constant	37.829	58.817	54.403	59.653
	(44.292)	(42.155)	(44.971)	(45.141)
Observations	37,124	37,124	37,124	37,124
R-squared	0.367	0.364	0.367	0.364
House Characteristics	Yes	Yes	Yes	Yes
Keywords	Yes	No	Yes	No
Transaction Characteristics	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Location FE	Yes	Yes	Yes	Yes
Agent FE	Yes	Yes	Yes	Yes
Cluster	ZIP Code	ZIP Code	ZIP Code	ZIP Code

Notes: Standard errors are clustered at the ZIP Code Level (*** p<0.01, ** p<0.05, * p<0.1).

Table 8: Description Uniqueness and Agent Experience

VARIABLES	(1) Ln(Price$_i$)	(2) Ln(Price$_i$)	(3) DOM	(4) DOM
$Unique_i$	0.687***	0.756***	28.835***	28.912***
	(0.180)	(0.216)	(10.011)	(10.409)
Total Sale #	-0.002**	-0.002**	0.081*	0.083*
	(0.001)	(0.001)	(0.048)	(0.046)
Total Sale # $\times Unique_i$	0.003***	0.003***	0.014	0.008
	(0.001)	(0.001)	(0.049)	(0.048)
Constant	11.487***	11.432***	47.905	52.900
	(0.212)	(0.230)	(43.278)	(44.535)
Observations	37,124	37,124	37,124	37,124
R-squared	0.887	0.878	0.372	0.368
House Characteristics	Yes	Yes	Yes	Yes
Keywords	Yes	No	Yes	No
Transaction Characteristics	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Location FE	Yes	Yes	Yes	Yes
Agent FE	Yes	Yes	Yes	Yes
Cluster	ZIP Code	ZIP Code	ZIP Code	ZIP Code

Notes: This table displays the estimation results of

$$Ln(Price_i) = \alpha + \theta_1 Unique_i + \theta_2 TotalSale + \theta_3 TotalSale \times Unique_i + X_i'\beta + \eta_z + \mu_c + \delta_t + \mu_c \times \delta_t + \varepsilon$$

Standard errors are clustered at the ZIP Code Level (*** p<0.01, ** p<0.05, * p<0.1).