International Journal of Strategic Property Management
ISSN: 1648-715X / eISSN: 1648-9179
2022 Volume 26 Issue 5: 345–361
https://doi.org/10.3846/ijspm.2022.17590

# DOES MACHINE LEARNING PREDICTION DAMPEN THE INFORMATION ASYMMETRY FOR NON-LOCAL INVESTORS?

Jinwoo JUNG, Jihwan KIM, Changha JIN*

*Department of Economics, College of Business and Economics, Hanyang University ERICA, Ansan, Republic of Korea*

Received 03 March 2022; accepted 30 August 2022

**Abstract.** In this study, we examine the prediction accuracy of machine learning methods to estimate commercial real estate transaction prices. Using machine learning methods, including Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machine (SVM), and Deep Neural Networks (DNN), we estimate the commercial real estate transaction price by comparing relative prediction accuracy. Data consist of 19,640 transaction-based office properties provided by Costar corresponding to the 2004–2017 period for 10 major U.S. CMSA (Consolidated Metropolitan Statistical Area). We conduct each machine learning method and compare the performance to identify a critical determinant model for each office market. Furthermore, we depict a partial dependence plot (PD) to verify the impact of research variables on predicted commercial office property value. In general, we expect that results from machine learning will provide a set of critical determinants to commercial office price with more predictive power overcoming the limitation of the traditional valuation model. The result for 10 CMSA will provide critical implications for the out-of-state investors to understand regional commercial real estate market.

**Keywords:** machine learning, office price, commercial real estate, prediction accuracy, information asymmetry, non-local investors.

## Introduction

The key role of markets is to coordinate mechanism that uses prices to convey information between stakeholders. Due to heterogeneities entrenched in each asset, geography, and transaction, however, the scarcity of commercial real estate price information makes the process of price discovery slower and the scope for asymmetrical information great (Geltner et al., 2003). This implies that the development of models that accurately measure commercial real estate prices assumes greater importance in the process of price discovery and eventually supports investment decisions.

The conventional approaches to measuring the price of a targeted property mainly consist of a sales comparison approach, a cost approach, and an income approach. First, a sales comparison approach has often been employed by investment practitioners and appraisers, as it actively takes into account the current market conditions. Despite its distinct benefit, due to careless selection of comparable properties and/or misspecification of weights given to each comparable, it may adjust sales prices inconsistently. Second, a cost approach is mainly utilized for new construction projects that may not have relevant comparisons. However, it is considered less reliable than sales comparison and income approaches, as it may not fully reflect demand conditions in the market. Lastly, an income approach that valuates income-generating properties via discounted cashflow (DCF) analysis and/or direct capitalization has also been utilized to determine market values. An income approach, however, has critical pitfalls to predict market values as well: it is often difficult to measure key elements of cashflows of leased properties, such as effective rental incomes, maintenance costs, and net operating incomes; furthermore, it is thorny to predict capitalization rates that reflect distinct features of a targeted property.

To address those issues and limitations, econometric techniques that rely on null hypothesis testing have commonly been employed to examine underlying dynamics in commercial real estate markets. Fundamentally, econometric techniques such as a hedonic approach can infer the value of real properties by examining property heterogeneities that consist of distinct structural, locational,

*Corresponding author. E-mail: *cjin@hanyang.ac.kr* and transaction characteristics. Due to the abundant availability and continuous quality improvement of information regarding commercial real estate markets in recent decades, those econometric techniques are increasingly able to address sampling errors and/or unobserved heterogeneity, so-called omitted variable bias (OVB). Furthermore, there have been numerous efforts to reflect distinct locational features by addressing spatial dependence and heterogeneity as well (McMillen & Redfearn, 2010; Parmeter et al., 2007).

These techniques, however, are still criticized, as they cannot fully reflect complex interrelationships between property values and their key determinants. According to Liu et al. (2015), for instance, non-local investors tend to overpay by an estimated 13.5% when they purchase and sell at an estimated 7% discount value relative to local investors. These empirical results imply that intermediation factors and/or transaction characteristics are also crucial determinants of property values.

The previous real estate literature suggests that such systematic differences between local and non-local investors may arise from selection bias, asymmetric information, and investor clienteles. The selection bias often occurs when non-local investors purchase extraordinary, incomparable properties such as prime offices (Liu et al., 2015). The asymmetric information that stems from intermediation factors and/or the lack of market knowledge will result in higher search costs and/or a higher variance in the distribution of their estimated price (Han & Hong, 2016; Turnbull & Sirmans, 1993). The investor clientele represents non-local investors who can systematically show premiums or discounts in commercial real estate transactions, for instance, due to their tax-exempt status. Furthermore, the previous literature commonly points out that such systematic distortions often arise from non-linear, complex relationships between agents (Han & Hong, 2016; Shi & Tapia, 2016; Turnbull & Sirmans, 1993). That is, standard parametric econometric approaches will reveal critical pitfalls to examine such complex relations due to their pre-built, rigid functional forms. As a result, novel approaches such as various machine learning (ML) techniques have become promising alternatives thanks to their flexible, evolving features to predict market dynamics (see e.g., Ho et al., 2021; Pérez-Rave et al., 2019; Simlai, 2021).

While conventional econometric approaches aim to infer causal relationships among correlated variables in the given data with pre-defined model specifications, machine learning aims to return the most accurate prediction model that is built over its complex learning process, led by artificial intelligence. In principle, machine learning keeps adjusting its "flexible" learning algorithms to improve prediction accuracy and model stability. Hence, we believe that machine learning approaches can alleviate concerns about the pitfalls of conventional models, which are grounded on pre-determined, inflexible statistical inference, and eventually support investment decisions by providing more accurate, reliable models for understanding the underlying mechanics of commercial real estate markets.

The goals of this study are as follows. First, this study aims to construct optimal, accurate predictive models to estimate commercial real estate transaction prices. We build predictive models based upon various machine learning approaches such as a Random Forest (RF), a Gradient Boosting Machine (GBM), a Support Vector Machine (SVM), and Deep Neural Networks (DNN), as well as conventional econometric approaches such as a standard parametric hedonic approach. We then focus on improving the predictive power by comparing the Root Mean Squared Error (RMSE), the Mean Absolute Percentage Error (MAPE), and the Mean Absolute Error (MAE) of each model, given the default and optimal settings of hyperparameters of each ML approach. By doing so, we eventually specify the best predictive model based upon the comparisons of the levels of prediction accuracy of each model, which are particularly differentiated based upon the nature of irregular transaction frequency, which is typically observed in commercial real estate trading.

Second, this study also aims to validate the power of flexible functional specifications in modeling commercial real estate prices. Like the previous real estate literature, we also posit that non-linear, complex relationships between variables can better explain the underlying dynamics in the commercial real estate market. Each ML method will reflect such non-linear relations between property values and key determinants by maintaining variables such as an investor status, a proxy for asymmetric information and/or investor clientele. As ML approaches outperform relative to standard "parametric" econometric approaches while reflecting such complex relations, we can conclude that ML methods reduce the generalization error of the prediction by addressing misspecification errors and/or entrenched non-linearities.

Lastly, this study also focuses on increasing the interpretability of ML models. One of the common criticisms regarding ML approaches mainly results from their complex algorithms and less clear functional specifications (Mullainathan & Spiess, 2017). Hence, this study provides a partial dependent plot (PD plot) that indicates the relative impact of each input feature on the predictability of each model. By doing so, this study can provide the influence of each feature on the predictive power of each model more clearly.

Correspondingly, based upon the 215,344 US office transactions data over the period of 2004–2017, we extend the literature on modeling commercial real estate prices by employing various ML frameworks. First, among various ML approaches and a standard hedonic approach, we find that both RF and GBM methods provide better predictive powers to estimate a commercial real estate transaction price at the national level. Furthermore, we categorize the sample into each CMSA level and provide the best predictive model for each CMSA. Second, ML methods successfully address misspecification errors and/or entrenched non-linearities relative to a standard hedonic approach while maintaining variables regarding investor status, which are representative features of such non-linearities. Like Liu et al. (2015), non-local investor behavior is a significant impact on transaction price in the hedonic based model due to an embedded misspecification problem associated with hedonic regression and controlling for all necessary variables. However, the influence of investor status in ML models is not considered as much critical a factor as in the regression model. Finally, we find that the partial dependent plot shows a less percentage of decreasing predictive power, and the relative decrease in accuracy of the ML model is marginal compared to the hedonic approach. Thus, the predictive estimate of transaction price will be a possible initial negotiating price as an unbiased initial reference price for non-local investors in the commercial real estate market.

In addition, the traditional valuation process may be biased by client feedback and the level of available information set to the commercial market (Hansz & Diaz, 2001; Gallimore & Wolverton, 2000). Especially, the real estate brokers and local appraisers are known to have an influential market information for their familiar market and are commissioned to overcome inferior information accessibility by uninformed market participants such as out-of-state investors. Thus, we expect out-of-state investors to employ these machine learning valuation models and to reduce a potential error caused by the traditional valuation method[1]. We believe the practitioners can leverage machine learning valuation to achieve more accurate investment value by collecting much larger amounts of available data from data providers. Therefore, this study provides a groundwork to practitioners to overcome a current challenged investment issues in commercial real estate transactions in a wide range of labor-intensive processes such as valuation, underwriting, portfolio valuation, assessments of underlying collateral, and potential risk management. With the increasing number of data providers and improved quality of available transaction data, we believe the commercial real estate investors will widely adopt the machine learning valuation method.

This paper is comprised of the following sections: Section 1 describes the literature review; Section 2 provides a description of the data; Section 3 outlines methodology; Section 4 presents an analysis of results; Section 5 presents a robustness check and, the last section presents a conclusion and policy implication.

## 1. Literature review

The development of models that accurately measure commercial real estate prices assumes greater importance in the process of price discovery and eventually supports investment decisions. In real estate and urban economic literature, a hedonic approach has generally been employed to measure prices and quantities in commercial real estate markets on account of the absence of asset fundamentals (Colwell et al., 1998; Hill, 2013). In principle, the hedonic relation arises due to heterogeneity: this model posits an explanation for a market containing heterogeneous properties, which possess diversified structural, spatial, and contract characteristics. By examining those heterogeneities, a standard parametric hedonic approach can infer values of real properties.

Appraisals are generally adopted the income approach using the capitalization of the net income of an asset observed from transactions in a similar real estate market. However, this approach is also criticized for using mismatched comparable in either time or building characteristics, and appraisers are also anchored on previous valuations or the previous transaction price of a building. Therefore, the estimated appraised price shows a lag of the market and provides smoothed approximations of true market prices (Kok et al., 2017).

In general, the hedonic approach is widely adopted in a real estate evaluation model, especially for mass appraisal. As the mass appraisal methods became more and more necessary, the International Association of Assessors (IAAO) proposed an automatic evaluation method (AVM) which suggested standardized methods such as Coefficient of Dispersion (COD) and Price-Related Differential (PRD) (International Association of Assessing Officers, 2013).

While hedonic AVMs are preferred for mass appraisals because of their simplicity based on simple regression models that are easy to implement and understand, AVMs are also criticized for inaccurate estimation caused by non-linear relationships between the predicted value and explanatory variables. Therefore, a single predictive formula model might not be successful at predicting property value most accurately.

This methodology often reveals critical pitfalls due to specification errors (Hill, 2013; McMillen & Redfearn, 2010) and/or sampling errors (Peterson & Flanagan, 2009). To mitigate those issues, non- or semi-parametric estimation techniques have consistently been employed. Due to the combination of highly flexible functional forms and spatially varying coefficients, these alternatives can estimate property values without imposing arbitrary contiguity matrices and/or distributional assumptions on the data (see e.g. McMillen & Redfearn, 2010; Meese & Wallace, 1991; Parmeter et al., 2007). In the same vein, machine learning approaches have increasingly played a

[1] The rise of machine learning allows a privately-held information main sourced from individual network available to reduce a disparity of information quality between of private network group and client who adopt machine learning based information platform (Egan, 2019). A major equity investment firm provides a value platform that uses machine learning tools to advise their clients. The new platform based on machine learning is now available for nonlocal investors and foreign client as investors with a tailored research reports for investors, and possible solution for upcoming strategic decision making for their portfolio including real estate. Thus, this will enhance information efficiency for nonlocal investor and foreign investors to dampen information asymmetry, allowing to negotiate with counterparts based on relatively accurate prediction on possible transaction price.

crucial role in inferring property values, as they can ideally reflect such non-linearities.

Since the traditional hedonic model is criticized for its statistical limitation and a researcher's quasi-selection bias toward a standard set of explanatory variables, we attempt to apply machine learning models which can improve prediction accuracy through numerous trials of combinations of unlimited explanatory variables, training, and testing the model on randomly selected parts of the datasets, leading to precise out-of-sample tests of predictive performance. The machine learning (ML) method in real estate valuation is suitable for conducting mass appraisal techniques because ML method will reflect such non-linear relations between property values and key determinants by maintaining variables and self-learning algorithms. A well-known case is the housing valuation model of Zestimate by the American agency Zillow (Kok et al., 2017).

Machine learning can be classified into two major categories: supervised learning and unsupervised learning (Bishop, 2006; Conway, 2018; Ho et al., 2021; Mullainathan & Spiess, 2017). Supervised learning is defined as the estimated relationship between a dependent variable and an observed outcome and applies it to new input to predict the new outcome. When the training data includes outcomes, it is referred as labeled data (Bishop, 2006; Mullainathan & Spiess, 2017). Unsupervised learning is defined as a method to uncover relationship among variables through hidden structures within given sample data.

The unsupervised learning prediction or estimation aims to cluster a set of variables and conducts an optimization process to provide predictive value. The K-mean clustering algorithm and artificial neural network are classified as unsupervised learning. The unsupervised learning is associated with clustering problems and dimensionality reduction problems. In this study, linear regression, Random Forest, Gradient Boosting Machines, Support Vector Machines, and Deep Neural Networks methods are classified into supervised learning to measure the prediction error on test data. Machine learning techniques are specifically appropriate for modeling complex hidden patterns and non-linearities that are entrenched in the relationships between property prices and their structural, spatial, and contract features (Cowden et al., 2019; Peterson & Flanagan, 2009). Unlike stochastic models, whose range of forecasting is often limited to the scope of variables they choose, furthermore, machine learning models can better predict a stochastic variety of commercial real estate prices without such limitations. Especially, we adopt the Neural Network method, which provides a practical alternative to the traditional least square model and efficiently analyzes the non-linearities in the underlying relationships among the parameters. We propose that neural networks could be robust to model misspecification and especially to various peculiarities in how various explanatory variables are measured (Peterson & Flanagan, 2020).

This study particularly focuses more on examining the impact of complex transaction processes (Wong et al., 2012; Zhou et al., 2015). Even though previous machine learning approaches have broadly been employed to address non-linearities that are embedded in structural and spatial features, the underlying dynamics within convoluted transactions have not been fully explored in the prediction of commercial real estate prices. By comparing the commercial real estate price prediction power of various machine learning techniques that include Random Forest (RF), Boosting, Support Vector Machines (SVM), and Deep Neural Networks (DNN), this study suggests implementable, reliable predictive models for commercial real estate prices.

## 2. Data

The raw data consist of 215,344 transaction-based office properties provided by Costar corresponding to the 2004–2017 period. However, we only include the office transaction data only for ten major CMSA; Boston, Chicago, Denver, Las Vegas, Los Angeles, Miami, New York, San Diego, San Francisco, and Washington DC. We also exclude data with one or more missing variables and property with special conditions such as auction, 1031 tax-deferred exchanges, and building contaminations. After verifying the data with regional matching and missing variables, we include a total of 19,640 transactions for ten major CMSA. Table 1 contains descriptive statistics for commercial property included in this study including transaction price, floor, building size(sf), land size(sf), age, number of parking space, number of tenants, CoStar 5 Star rating classification system[2], Broker information[3], Sunbelt as indicating variable if property located in Sunbelt[4], Investor Status (non-local buyer or seller), and CMSA GDP[5]. The average transacted property price is $ 1,483,627, which is 14.21 in natural logarithm value, and

[2] The CoStar 5 Star Building Rating System provides a national rating for commercial buildings. Properties are evaluated and rated using a universally recognized 5 Star scale based on the characteristics of each property type, including: architectural attributes, structural and systems specifications, amenities, site and landscaping treatments, third party certifications and detailed property type specifics (see more at https://www.costar.com/).

[3] Top global brokerage firm list obtained from reonomy website, https://www.reonomy.com/. This website provides top sixteenth CRE brokerage companies lists; Cushman and Wakefield, CBRE, SVN, Lee & Associates, JLL, Colliers, NAI Global, Avison Young, Transwestern, Marcus & Millichap, Kidder Matthews, Newmark Knight. Frank, RE/MAX Commercial, Keller Williams, Savills, Coldwell Banker.

[4] The sunbelt is a region of the US generally considered to stretch across the Southeast and Southwest. The sunbelt consists of 13 states; Alabama, Arizona, California, Florida, Georgia, Louisiana, Mississippi, Nevada, New Mexico, North Carolina, South Carolina, Texas, Utah.

[5] Gross domestic product estimates the value of the goods and services produced in each CMSA. A comprehensive measure of the economies of each CMSA areas (https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas).

Table 1. Descriptive statistics

| Variable | Mean | S.D. | Min. | Max. |
|---|---|---|---|---|
| Transaction price (ln) | 14.21 | 1.56 | 2.30 | 20.99 |
| Assessed value (ln) | 13.57 | 1.99 | 3.00 | 26.89 |
| Building SF (ln) | 9.64 | 1.32 | 6.14 | 15.17 |
| Land SF (ln) | 10.29 | 1.35 | 4.23 | 17.27 |
| Building Age | 43.74 | 27.73 | 1.00 | 288.00 |
| Building Age$^2$ | 2,681.81 | 4,092.27 | 1.00 | 82,944.00 |
| Floor | 2.88 | 4.22 | 1.00 | 110.00 |
| # of Parking | 101.69 | 196.20 | 1.00 | 4,051.00 |
| # of Tenant | 6.97 | 11.05 | 1.00 | 226.00 |
| CoStar Rating | 2.49 | 0.72 | 1.00 | 5.00 |
| Buyer Broker | 0.13 | 0.33 | 0.00 | 1.00 |
| Seller Broker | 0.24 | 0.43 | 0.00 | 1.00 |
| Nonlocal Buyer | 0.21 | 0.40 | 0.00 | 1.00 |
| Nonlocal Seller | 0.43 | 0.49 | 0.00 | 1.00 |
| Sunbelt | 0.57 | 0.50 | 0.00 | 1.00 |
| CMSA GDP (ln) | 13.13 | 0.80 | 11.40 | 14.35 |

*Note:* This table presents summary statistics for total 19,640 office property transactions for 10 major CMSA The first column shows name of each variable for analysis. Mean is sample mean. S.D. is standard deviation. Min and Max are minimum and maximum, respectively. The transaction price is denoted as price (ln) which is a natural logarithm of the transaction sale price in million U.S. dollars. The assessed value is a natural logarithm of the assessed value of property in million U.S. dollars. Building SF (ln) is a natural logarithm of rentable building area, measured in square foot. Land SF (ln) is a natural logarithm of the gross square foot of the lot. Building Age represents age of building at the sale date. Building Age$^2$ present a possible U-shape effect from development effect. Floor is the total number of floors in the office building. # of parking is the total number of parking lots in the office building. # of Tenant is the total number of persons in the office building. CoStar Rating is a new CoStar's five-star building rating system that replaces the existing classification system. Buyer Broker is a one if buyer's broker is listed in top 16$^{th}$ global brokerage firms. Broker (Seller) is a one if seller's broker is listed in top global brokerage firms. Non-local Buyer and Non-local Seller are also indicator variables, taking one value of one if the buyer's address is a different geographic state. Sunbelt is indicating variables, taking on a value of one when the office building is located in the Sunbelt State. CMSA GDP (ln) is a natural logarithm of gross domestic product for 10 major CMSA in million U.S. dollars.

the assessed value is $ 782,305, which is 13.57 in natural logarithm value. The property size(ln) is 15,367 SF which is 9.64 in natural logarithm value. The average building age is approximately 43 years, and the number of floors is 2.88 floors. The average number of parking is 101.69, and the average number of tenants is approximately 7 tenants in each building. Also, approximately 13% of buyer's broker is listed within the top 16th broker company, and 24% of seller's broker is one of top 16th broker company. A total 21% of transactions occurred in the Sunbelt state. A total 21% of buyer is non-local buyer and 43% of seller is non-local seller involved in transaction, respectively.

It is worth noting that, unlike capital assets such as stocks and bonds that are traded frequently, commercial real estate transactions irregularly occur in nature, and so the total observations in our study are also irregularly occurring series of office transactions over the study period. We postulate that this nature of irregular data frequency that is typically embedded in real estate trading will decisively have effects on the learning process and, eventually, the prediction power of each machine learning approach.

## 3. Methods

### 3.1. Standard parametric hedonic methods

The hedonic approach decomposes expenditures on commercial real estate into measurable prices and quantities so that prices for different assets or for identical assets in different places can be predicted and compared (Hill, 2013; Meese & Wallace, 1991). Existing hedonic models have employed various functional forms, and the advantages and disadvantages of each model substantially rely on their functional forms. A standard parametric hedonic approach, for instance, is a simple way to measure price. Due to its rigid structure, however, this method commonly reveals omitted variable bias, which refers to bias that arises from missing characteristics of assets, and inflexibility that exerts severe restrictions on the potential interactions between features (Hill, 2013; Peterson & Flanagan, 2009).

To examine discrepancies between a conventional hedonic approach and various machine learning approaches more explicitly, this study constructs a hedonic price model as follows:

$$\ln\left(P\right) = X\gamma_1 + Z\gamma_2 + T\gamma_3 + G\gamma_4 + \varepsilon, \tag{1}$$

where: $\ln(P)$ denotes the log of the transaction price; $X$ denotes a matrix of property characteristics; $Z$ denotes a matrix of transaction characteristics; $T$ denotes a matrix of time period dummies; $G$ denotes a matrix of geographic dummies; $\varepsilon$ denotes the error term.

### 3.2. Ensemble methods

The Classification and Regression Tree (CART) is a tree-based decision process based upon the recursive partitioning algorithm, characterized as its intuitive explanatory power to examine non-linear relations and flexibility to support multi-class regression and classification (Čeh et al., 2018). As a result, the CART has incrementally been employed to predict commercial and residential real estate prices (Čeh et al., 2018; Gupta et al., 2022; Ho et al., 2021; Pérez-Rave et al., 2019; Yilmazer & Kocaman, 2020).

The predictive performance of CART models can be improved specifically through various ensemble learning techniques, such as bagging, boosting, and stacking (Breiman et al., 2017). An ensemble learning technique refers to a process that aggregates the prediction of distinct, diverse models, which stem from different modeling algorithms and/or different training data sets. That is, an ensemble model performs as a single model, even though it works with multiple base models within the model. By doing so, this approach aims to reduce the generalization error of the prediction, which consists of bias, variance, and irreducible errors (Kotu & Deshpande, 2014).

According to the type of prediction error in which ensemble learning techniques specifically address, ensemble methods can broadly be categorized into two groups: bagging and boosting. The variance of the prediction error commonly arises when models are overfitting on a specific given number of training data points. As Breiman (1996) suggested, bagging can better perform and reduce the variance of classification and regression trees by aggregating individual base models that have distinct statistical properties. The bias of the prediction error, on the other hand, typically arises when models are not learning enough from the training data and hence lead to unreliable predictions and poor generalization. The boosting algorithm converts a collection of weak learners into accurate, reliable learners (Schapire, 1990; Schapire & Freund, 1995).

According to ways of how models generate learners, furthermore, ensemble methods can broadly fall into two groups: parallel and sequential ensemble techniques. Parallel ensemble methods basically generate base learners in a parallel format to secure independence between the base learners. Random Forest models, for instance, which is a tweaked version of bagging, mitigate the variance of the prediction error by combining weak base learners that are independent to form a single strong learner (Breiman, 2001). Due to their ensemble procedures, however, parallel ensemble methods particularly perform less with small training datasets.

Sequential ensemble methods improve the predictive performance of models by assigning higher weights to previously misrepresented learners in which data dependency resides. Gradient boosting, for example, reduces the bias of the prediction error, as new predictors are fit to counter the impacts of prediction error in the preceding predictors (Friedman, 2001). When the depth of CARTs and/or the number of boosting iterations increase, however, sequential ensemble methods can overfit the training data. To prevent this undesirable pitfall, sequential ensemble methods regularize their procedures by modifying the update rule, penalizing the complexity of trees, and employing stochastic boosting (Friedman, 2002; Telgarsky, 2013).

### 3.3. Random Forest (RF)

To examine distinct features of ensemble methods explicitly, this study employs Random Forest and Gradient Boosting methods to predict commercial real estate prices, respectively. In our Random Forest (RF) models, the prediction of $B$th tree for an input vector $X$ can be represented by:

$$\hat{f}^{\,Random\,Forest}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{\,tree}(X \mid b). \tag{2}$$

In the function estimation problem, meanwhile, models aim to find a function $F^{*}(X)$ that maps a set of random input $X$ to a random output $y$, given a training data set $\{y_i, X_i\}_1^N$ over the joint distribution of all $(y, X)$ values:

$$F^{*}(X) = \arg\min_{F(X)} E_{y,X}\psi(y, F(X)), \tag{3}$$

where the expected value of some specified loss function $\psi$ is minimized.

### 3.4. Gradient Boosting Machines (GBM)

Boosting approximates $F^{*}(X)$ to the weighted sum of functions from weak learners:

$$(\beta_m, \boldsymbol{a}_m) = \arg\min_{\beta,\boldsymbol{a}} \sum_{i-1}^{N}\psi\left(y_i, F_{m-1}(X_i) + \beta h(X_i; \boldsymbol{a})\right), \tag{4}$$

where: $h(X;\boldsymbol{a})$ denotes a base learner; $\boldsymbol{a}$ denotes the parameters; $\beta$ denotes the expansion coefficients; $m$ denotes boosting iterations. According to Friedman (2001), given $h(X;\boldsymbol{a}_m)$, gradient boosting approximates the equation (NUMBER) for arbitrary loss functions $\psi$, and the optimal value of the coefficient $\beta_m$ is determined as follows:

$$\beta_m = \arg\min_{\beta}\sum_{i=1}^{N}\psi\left(y_i, F_{m-1}(X_i) + \beta h(X_i; a_m)\right). \tag{5}$$

### 3.5. Support Vector Machines (SVM)

In the early 1990s, Vapnik (2013) suggested an algorithm for classification that has consistently evolved into Support Vector Machines (SVM), which has extensively been employed for performing data classification and prediction in a supervised machine learning framework. SVM typically utilizes linear and non-linear separating planes – such as linear and polynomial kernel functions, radial basis functions (RBF), neural networks, and multi-dimensional splines – to train classifiers (Cortes & Vapnik, 1995; Izmailov et al., 2013). Those transformers turn the original space into a multi-dimensional feature space and eventually allow these separating planes to have a maximal margin to contain all the points with a very small error. Due to this distinct feature, in real estate research, SVM has increasingly been employed to predict real estate prices (Ho et al., 2021; Lam et al., 2009), to predict default (Cowden et al., 2019), and to analyze developers' decision making (Rafiei & Adeli, 2016).

The main limitation of SVM techniques for predicting real estate prices is that they are black-box models, which are not able to have clear functional specifications. Furthermore, SVM is fundamentally designed to support binary classification. To mitigate this limitation, however, extended models and frameworks have consistently been developed, such as multi-class SVM, one-hot encoding for categorical features, and Support Vector Regression (SVR) algorithms for working with continuous and categorical features (see e.g. Ho et al., 2021; Rafiei & Adeli, 2016).

This study employs a SVM approach, which inputs the vector of explanatory variables for *i* th office transaction, $X_i$, that will be mapped onto a higher-dimensional feature space to optimize the problem as follows:

$$minimize \frac{1}{2} W^T W + C \sum_{i=1}^{N} u_i^2, s.t. \gamma_i \left( W^T \phi\left( X_i \right) + b \right) \geq 1 - u_i,$$
$$i = 1, \ldots, N \qquad (6)$$

where: $N$ denotes the number of transactions; $W$ denotes the parameter vector of $X_i$; $C$ denotes a parameter that normalizes the error terms $u_i^2$; $\phi\left(X\right)$ denotes the kernel function for mapping the features; $b$ denotes the intercept. We can state a positive semidefinite symmetric function, $K\left(x, \gamma\right)$, as an inner product of $\phi\left(x\right)$ and $\phi\left(\gamma\right)$:

$$K\left(x,\gamma\right) = \left\langle \phi\left(x\right), \phi\left(\gamma\right) \right\rangle,$$
$$if\ and\ only\ if \int K\left(x,\gamma\right) g\left(x\right) g\left(\gamma\right) dx d\gamma \geq 0 \, \forall g. \qquad (7)$$

By taking the inner product of $\phi$, the new feature space can explicitly be found. Using the RBF kernel, for example, a grid search over parameters such as $C$ and $\gamma$ are performed to determine the highest predictive power on the test data set.

### 3.6. Deep Neural Networks (DNN)

Over the past decade, Deep Neural Networks (DNN) have increasingly played a vital role in the field of real estate research. DNN refers to computing systems that can be characterized as Artificial Neural Networks (ANNs), particularly associated with the use of "multiple" hidden layers within the network. Due to this distinct feature, deep learning architectures and their learning algorithms have the capacity to model complex non-linear relationships. Hence, DNN models have broadly been adopted to predict real estate prices (Nghiep & Al, 2001; Xu & Gade, 2017), to assess real estate values (Shen & Ross, 2021; Yu et al., 2018; You et al., 2017), and to examine urban dynamics (Yao et al., 2021).

Despite DNNs' abundant benefits that model complex non-linearities, deep learning algorithms commonly confront the following pitfalls. First, DNN models are vulnerable to overfitting and high variance when the abstraction layers empower them to model sparse dependency in the training data set (Rice et al., 2020; Sun et al., 2017). To address this issue, it is essentially required to increase the number of training data sets and/or to employ regularization techniques such asdropout and weight decay. Second, DNN models can encounter vanishing gradient problems when they utilize a back-propagation learning algorithm for updating network weights to minimize error (Hochreiter, 1998). To mitigate this issue, gating mechanisms, such as Gated Recurrent Unit (GRU), and tweaked learning mechanisms, such as Long Short-Term Memory (LSTM), are increasingly utilized.

## 4. Results

In Table 2, we conduct a standard hedonic model to examine whether investors choosing a particular property have a marginal willingness to pay for an attribute equal to the derivative of the hedonic price function. In addition, we also consider a possible premium or discount in the transaction associated with non-local buyer and seller (Liu et al., 2015; Ling et al., 2018; Kandlbinder et al., 2018). The traditional hedonic model can be written as follows:

$$\ln\left(P_i\right) = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \gamma_3 NL_i + \gamma_4 Time + \gamma_5 \, Region + \varepsilon_i. \qquad (8)$$

The log of the transaction price of the office ln*($P_i$)* is a function of office property (*i*), observable individual office property characteristics ($X_i$), CoStar 5 star rating as $Z_i$, $NL_i$ as an indicator variable based on investor status such as non-local buyer or seller, and 2×2 indicating variable that represents buyer and/or seller status as if they are non-local buyers or sellers. *Time* and *Region* is indicator controlling for time and region[6] fixed effect. And, $\varepsilon_{iz}$ represents the error term. The advantage of this specification is that we can examine the theory of whether non-local buyer and seller pay premiums or discount in commercial real estate transactions.

We formally test for determinants for office transaction price in the hedonic model associated with investor status. We find positive premiums on assessed value, building size, building age, floor, number of parking, number of tenants, and CoStar rating. We find the use of top brokerage services increases both buyer and seller's transaction prices which may relate to higher quality selection bias by top listing brokerage firms. In models (1)

[6] We control for region at CMSA level in addition to sunbelt in current result. The current analysis adopts transaction-based data and thus there may be higher frequency of transaction observation if investors are more interested in. Therefore, we believe that the current result can reflect the investment behavior in investment market with relatively high observed transaction in preferred market by investor at the CMSA level.

Table 2. Hedonic regression analysis for the office transaction price

| Dependent variables | Transaction Price (ln) | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Assessed Value | 0.467$^{***}$ | 0.476$^{***}$ | 0.471$^{***}$ | 0.473$^{***}$ | 0.473$^{***}$ | 0.467$^{***}$ |
| | (0.004) | (0.004) | (0.004) | (0.004) | (0.004) | (0.004) |
| Building SF | 0.287$^{***}$ | 0.292$^{***}$ | 0.290$^{***}$ | 0.289$^{***}$ | 0.292$^{***}$ | 0.287$^{***}$ |
| | (0.010) | (0.010) | (0.010) | (0.010) | (0.010) | (0.010) |
| Land SF | –0.084$^{***}$ | –0.086$^{***}$ | –0.085$^{***}$ | –0.086$^{***}$ | –0.086$^{***}$ | –0.084$^{***}$ |
| | (0.007) | (0.007) | (0.007) | (0.007) | (0.007) | (0.007) |
| Building Age | 0.012$^{***}$ | 0.012$^{***}$ | 0.012$^{***}$ | 0.012$^{***}$ | 0.012$^{***}$ | 0.012$^{***}$ |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Building Age$^2$ | –0.0001$^{***}$ | –0.0001$^{***}$ | –0.0001$^{***}$ | –0.0001$^{***}$ | –0.0001$^{***}$ | –0.0001$^{***}$ |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Floor | 0.024$^{***}$ | 0.024$^{***}$ | 0.024$^{***}$ | 0.024$^{***}$ | 0.023$^{***}$ | 0.024$^{***}$ |
| | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| Parking | 0.0004$^{***}$ | 0.0004$^{***}$ | 0.0004$^{***}$ | 0.0004$^{***}$ | 0.0004$^{***}$ | 0.0004$^{***}$ |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Tenant | –0.014$^{***}$ | –0.014$^{***}$ | –0.014$^{***}$ | –0.014$^{***}$ | –0.014$^{***}$ | –0.014$^{***}$ |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Rating2 Star | 0.112$^{***}$ | 0.104$^{***}$ | 0.106$^{***}$ | 0.107$^{***}$ | 0.107$^{***}$ | 0.111$^{***}$ |
| | (0.032) | (0.032) | (0.032) | (0.032) | (0.032) | (0.032) |
| Rating3 Star | 0.252$^{***}$ | 0.245$^{***}$ | 0.246$^{***}$ | 0.249$^{***}$ | 0.248$^{***}$ | 0.252$^{***}$ |
| | (0.036) | (0.036) | (0.036) | (0.036) | (0.036) | (0.036) |
| Rating4 Star | 0.828$^{***}$ | 0.846$^{***}$ | 0.841$^{***}$ | 0.836$^{***}$ | 0.841$^{***}$ | 0.829$^{***}$ |
| | (0.048) | (0.048) | (0.048) | (0.048) | (0.048) | (0.048) |
| Rating5 Star | 1.784$^{***}$ | 1.819$^{***}$ | 1.803$^{***}$ | 1.796$^{***}$ | 1.818$^{***}$ | 1.784$^{***}$ |
| | (0.103) | (0.103) | (0.103) | (0.103) | (0.103) | (0.103) |
| Buyer_Broker | 0.127$^{***}$ | 0.127$^{***}$ | 0.130$^{***}$ | 0.121$^{***}$ | 0.130$^{***}$ | 0.127$^{***}$ |
| | (0.021) | (0.021) | (0.021) | (0.021) | (0.021) | (0.021) |
| Seller_Broker | 0.093$^{***}$ | 0.101$^{***}$ | 0.097$^{***}$ | 0.096$^{***}$ | 0.101$^{***}$ | 0.093$^{***}$ |
| | (0.016) | (0.016) | (0.016) | (0.016) | (0.016) | (0.016) |
| Nonlocal_Buyer | 0.202$^{***}$ | | | | | |
| | (0.017) | | | | | |
| Nonlocal_Seller | –0.106$^{***}$ | | | | | |
| | (0.013) | | | | | |
| Local Buyer & Local Seller | | 0.016 | | | | |
| | | (0.013) | | | | |
| Non-local Buyer & Local Seller | | | 0.222$^{***}$ | | | 0.211$^{***}$ |
| | | | (0.023) | | | (0.024) |
| Local Buyer & Non-local Seller | | | | –0.141$^{***}$ | | –0.102$^{***}$ |
| | | | | (0.014) | | (0.015) |
| Non-local Buyer & Non-local Seller | | | | | 0.094$^{***}$ | 0.091$^{***}$ |
| | | | | | (0.021) | (0.022) |
| Sunbelt | 0.551$^{***}$ | 0.529$^{***}$ | 0.533$^{***}$ | 0.529$^{***}$ | 0.555$^{***}$ | 0.550$^{***}$ |
| | (0.041) | (0.041) | (0.041) | (0.041) | (0.041) | (0.041) |
| CMSA GDP | 1.914$^{***}$ | 1.932$^{***}$ | 1.890$^{***}$ | 1.917$^{***}$ | 1.970$^{***}$ | 1.911$^{***}$ |
| | (0.212) | (0.213) | (0.213) | (0.213) | (0.213) | (0.212) |
| Year control | Yes | Yes | Yes | Yes | Yes | Yes |
| Region control | Yes | Yes | Yes | Yes | Yes | Yes |
| Constant | –19.751$^{***}$ | –20.144$^{***}$ | –19.533$^{***}$ | –19.819$^{***}$ | –20.627$^{***}$ | –19.712$^{***}$ |
| | (2.788) | (2.799) | (2.793) | (2.792) | (2.799) | (2.789) |
| Observations | 19,640 | 19,640 | 19,640 | 19,640 | 19,640 | 19,640 |
| $R^2$ | 0.681 | 0.678 | 0.680 | 0.680 | 0.679 | 0.681 |
| Adjusted $R^2$ | 0.681 | 0.678 | 0.679 | 0.679 | 0.678 | 0.681 |
| Residual Std. Error | 0.882 | 0.886 | 0.884 | 0.884 | 0.885 | 0.882 |
| F Statistic | 1,074.850$^{***}$ | 1,088.538$^{***}$ | 1,095.892$^{***}$ | 1,096.705$^{***}$ | 1,090.089$^{***}$ | 1,047.950$^{***}$ |

*Note:* This table presents regression analysis for a total 19,640 office property transactions for 10 major CMSA. We also control for time and region fixed effect. The significance p-value levels at 10%, 5%, and 1% is denoted as *, **, and ***, respectively.

and (2), we find that while the non-local buyer pays a premium on office property transaction compared to that of local buyer when they purchase, non-local seller transacted an office property with a discount compared to local seller when they sell an office property. We further classify the transaction into 2×2 cases such a transaction between a local buyer and local seller, between non-local buyer and local seller, between local buyer and non-local seller, and between non-local buyer and non-local seller transaction. The transaction between local buyer and local seller shows a slightly higher transaction price compared to other transactions. The transaction between non-local buyer and local seller shows a higher price than other transaction groups. The transaction between non-local seller and local buyer shows a less transaction price compared to other transactions. The transaction between non-local buyer and non-local seller shows 8.7% of the transaction premium, but this premium is lower than the transaction non-local buyer to local seller, which is 22.7% of the premium on property price. In general, the results strongly support the research hypothesis that non-local buyers pay a premium on office transactions and seller pay a discount. We attribute this discount and premium to asymmetric information and increased search costs that appear to reduce the negotiation control in office transactions. In Table 3, we present the optimization assumption of ML hyperparameter and calculate ranges for all hyperparameters. In order to estimate optimal defaults of ML algorithms, the default we follow the optimization process by Probst et al. (2019). Table 1 provide a definition of hyperparameter and optimal and default value of ML algorithms including Random Forest, Gradient Boosting Machines, Support Vector Machines, and Deep Neural Networks. We attempt to measure the tunability of the applicable algorithm and define hyperparameters to verify the difference between the default setting of hyperparameters and the optimal setting of hyperparameters. We also interpret the tunability value of individual parameters as how much performance can be improved by tuning each hyperparameter. The optimal value for each ML model is described in column.

In the first model, the random forest has two hyperparameter values; Number of trees and Max depth. The number of trees in Random Forest algorithm is nothing but a question of how many trees we should consider. Thus, the number of trees means a number of uncorrelated trees we ensemble to create the random forest. The Max depth of a tree in Random Forest is defined as the longest path between the root node and the leaf node. Random forest and gradient boosting machines are similar in basis of a large number of trees. While random forest approach follows the 'average or majority rules' and combines at the end of the process, gradient boosting machine is also combining decision trees but it starts the combining process at the beginning instead of at the end. The Min-rows specify the minimum number of observations for node size. Learn rate is defined as a weighting factor for correction by new trees added to the existing model when new

Table 3. Default and optimization of ML hyperparameter

| Hyperparameters of model | Default hyperparameter | Range of hyperparameters e | Optimal hyperparameter |
|---|---|---|---|
| 1. Random Forest (RF) | | | |
| Number of trees (ntrees) | 300 | 500, 1000, 2000 | 1000 |
| Max-depth | 10 | 30, 40, 50 | 50 |
| 2. Gradient Boosting Machines (GBM) | | | |
| Number of trees (ntrees) | 500 | 1000, 1500, 2000 | 2000 |
| Max-depth | 20 | 5, 10, 30 | 30 |
| Min-rows | 0 | 5, 10, 30 | 5 |
| Learn-rate | 0.01 | 0.1, 0.01 | 0.1 |
| Sample-rate | 0.8 | 0.5, 0.95 | 0.95 |
| 3. Support Vector Machines (SVM) | | | |
| Epsilon (ε) | 0.1 | 0.05, 0.2 | 0.1 |
| Cost (*c*) | 0.01 | 0.01, 0.003, … 14.4 | 0.994 |
| Sigma (σ) | 0 | 0.000, … 0.835 | 0.021 |
| 4. Deep Neural Networks (DNN) | | | |
| Hidden layer | 32 | 32, 64, 128 | 64, 64 |
| Learn-rate | 0.01 | 0.01, 0.02 | 0.02 |
| Learning rate annealing | 1.0E-6 | 1.0E-6, 1.0E-7, 1.0E-8 | 1.0E-7 |

*Note:* In order to estimate optimal defaults of ML algorithms, the default we follow the optimization process by Probst et al. (2019). Table 1 provides a definition of hyperparameter and optimal and default value of ML algorithms including Random Forest, Gradient Boosting Machines, Support Vector Machines, and Deep Neural Networks. They suggest measure for estimating the tunability of the applicable algorithm and define hyperparameters to verify the difference between the default setting of hyperparameters and the optimal setting of hyperparameters.

Table 4. Results of RMSE as a measure of prediction accuracy for whole and 10 CMSA

| | Number of observations | OLS | RF | GBM | SVM | DNN |
|---|---|---|---|---|---|---|
| Single CMSA level | | | | | | |
| Boston | 1,167 | 0.751 | *0.671* | 0.723 | 1.490 | 0.823 |
| Chicago | 1,041 | 0.893 | *0.762* | 0.784 | 1.840 | 1.075 |
| Denver | 1,289 | 0.744 | *0.613* | 0.685 | 1.050 | 0.898 |
| Las Vegas | 1,081 | 0.618 | 0.671 | *0.552* | 0.895 | 0.627 |
| Los Angeles | 5,668 | 0.728 | *0.481* | 0.497 | 0.778 | 0.707 |
| Miami | 753 | 0.511 | 0.419 | *0.406* | 1.140 | 0.471 |
| New York | 3,276 | 1.050 | *0.780* | 0.806 | 1.070 | 1.029 |
| San Diego | 1,587 | 0.708 | 0.481 | *0.469* | 0.844 | 0.608 |
| San Francisco | 2,050 | 0.663 | 0.922 | *0.545* | 0.985 | 0.654 |
| Washington D.C. | 1,728 | 0.644 | 1.037 | *0.578* | 1.340 | 0.718 |
| 10 CMSA level | | | | | | |
| RMSE | 19,640 | 0.881 | *0.633* | 0.645 | 0.756 | 0.880 |
| MAE | 19,640 | 0.643 | 0.384 | *0.367* | 0.485 | 0.594 |
| MAPE | 19,640 | 4.691 | 2.826 | *2.678* | 3.568 | 4.456 |
| $R^2$ | 19,640 | 0.681 | *0.842* | 0.836 | 0.776 | 0.695 |
| Rank | – | 5 | 1 | 2 | 3 | 4 |

*Note:* This table presents performance of each predictive model; OLS, Random Forest (RF), Gradient Boosting Machines (GBM), Support Vector Machines (SVM), and Deep Neural Networks (DNN). The Root Mean Square Error (RMSE) is the standard deviation of the prediction errors. Since the errors are squared before the errors are averaged, the RMSE gives a relatively high weight to large errors. Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of forecasts without considering their direction. The MAE measures absolute average differences that are weighted equally in the predictive value. The Mean Absolute Percentage Error (MAPE) is the mean or average of the absolute percentage errors of forecasts. Error is defined as actual or observed value minus the forecasted value, and percentage errors are summed without regard to sign to compute MAPE. $R^2$ is the percentage of the dependent variable variation that a predictive model explains and measures the scatter of the sample observation around a predictive model. For the same data set, higher $R^2$ measure represents smaller differences between the observed data and the predictive values. The analysis is based on a total of 19,640 office property transactions for 10 major CMSA. Thus, the result represents a predictive model for a commercial real estate market representing whole 10 CMSAs and each single CMSA.

trees are created to correct the residual errors in the predictions from the existing sequence of trees. A technique to slow down the learning in the gradient boosting model is to apply a weighting factor for the corrections by new trees when added to the model. This weighting is called the shrinkage factor or the learning rate, which specifies the learning rate. The range is 0.0 to 1.0, and the default value is 0.1. Sample rate defines the row (x-axis) sampling rate (without replacement). The sampling can improve generalization and lead to lower validation and test set errors. In Support Vector Model, the Epsilon ($\varepsilon$) specifies the epsilon-tube within which no penalty is associated with the training loss function with points predicted within a distance epsilon ($\varepsilon$) from the actual value. Cost ($c$) is referred to when the optimization problem to optimize both the fit of the line to data penalizing the amount of samples inside the margin at the same time, where cost ($c$) defines the weight of how much samples inside the margin contribute to the overall error. Consequently, with a low cost, samples inside the margins are penalized less than with a higher cost. Sigma ($\sigma$) determines how fast the similarity metrics decrease as parameterize the Gaussian kernel used to estimate non-linear classification. In Deep Neural Networks (DNN), a hidden layer is referring a layer located between the input and output of the algorithm, in which the function applies weights to the inputs and directs them through an activation function as the output. Similar to that of GBM, learn rate is a hyperparameter that allows how much correction of the model to apply to the existing model in response to the estimated error each time the model weights are updated. Learning rate annealing is a parameter that determines the schedule for learning rates starting with a relatively high learning rate and then lowering the learning rate during training. The first column includes the hyperparameters of the model, and the second column presents the default hyperparameter. The rangers of hyperparameter and optimal hyperparameter value are included in columns 3 and 4.

Table 4 represents the performance of the predictive model for single CMSA and whole CMSA level with a transaction price as predictive value and a set of explanatory variables in Eq. (1) and Table 2. In consistent with previous literature on performance measures on market predictability, we also adopt the Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and $R^2$ as a critical performance measure[7]. In our whole sample model, we find that the Random Forest Model shows a superior prediction accuracy against Gradient Boosting Machine, Support Vector Machine, Deep Neural Networks, and Hedonic Price Model. Compared to other predictive models, Random Forest and Gradient Boosting Machine belongings to ensemble tree model show its RMSE measure of 0.633 and 0.645, respectively, and a better prediction accuracy than Support Vector Machines, and Deep Neural Networks, 0.765 and 0.880 as RMSE measure. As expected, the Ordinary Least Square shows the highest inaccuracy with 0.881 as RMSE measure. We attribute a better performance of Random Forest and Gradient Boosting Machine to the bagging algorithm process and boosting algorithm process to provide a best fitting set of estimates to predict the test data in sample transactions[8]. In addition, the Random Forest and Gradient Boosting Machine method is known as an advantageous predictive model to predict the problems of multicollinearity issues among data and is free from the issue of missing variables. While Support Vector Machine is advantageous for predicting unstructured and semi-structured data and efficiently solve complex issues with appropriate kernel function, it is less efficient for large data sets (Ho et al., 2021)[9].

Table 4 also represents the results of prediction accuracy for each CMSA level[10]. For each CMAs, the top three CMAs (Los Angeles, New York, and San Francisco) account for 55.97% of the whole sample observations, while the top five (Los Angeles, New York, San Francisco, Washington D.C., San Diego) account for 72.86% of the sample, respectively. Similar to the result of whole 10 CMAs samples, Random Forest performed best in Boston, Chicago, Denver, Los Angeles, and New York. Since Random Forests are based on bagging algorithms that aim to control overfitting and reduce variance with independent classifiers, thus it is meaningful to predict the transaction price in the above city as independent classifiers. Similarly, those CMSAs such as Las Vegas, Miami, San Diego, San Francisco, and Washington D.C. shows Gradient Boosting Machine as best performed model. Likewise, to predict the transaction price accurately, we should understand the boosting algorithms, which is an approach to reduce bias and variance based on sequential classifiers.

While it is important to know the model accuracy to predict the transaction price, we believe it is also worthy to interpret the relative importance of each research variable to each model. The purpose of our research is not only to find the best accuracy model but also to identify which independent variables are most important to predict the model. Therefore, we conduct an analysis to evaluate the importance of a variable for predicting transaction price suggested by Breiman (2001)[11]. The advantage of relative variable importance is to enhance the interpretability of the predictive model. In a practical sense, it is also important to understand the reason behind the best predictive commercial property transaction

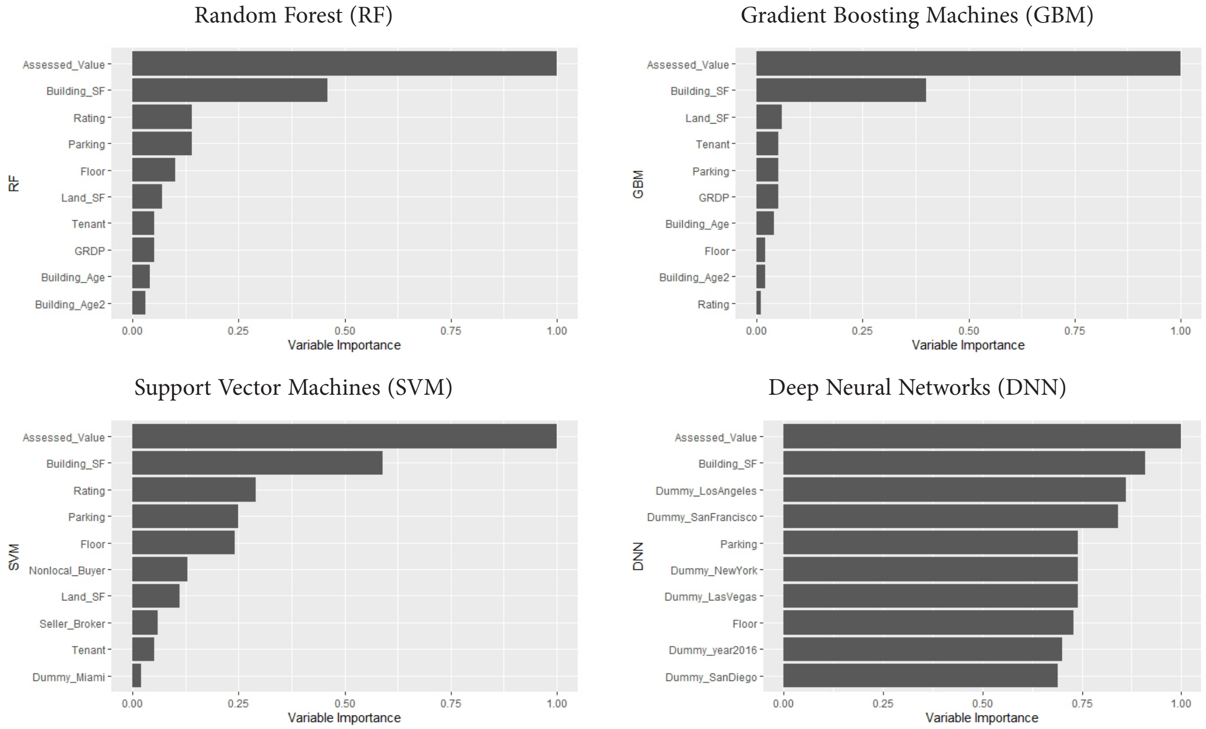
[7] RMSE (Root Mean Square Error) is the standard deviation of the prediction errors and defined as follows; $RMSE=\sqrt{\frac{\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2}{n}}$ where $y_i$ represents the predicted value of transaction price; $\hat{y}_i$ represents the actual of transaction price; where $\overline{y}_i$ represents the mean value of transaction price; and $n$ represents the number of observation in the test data. MAE (Mean Absolute Error) measures the average magnitude of the errors and defined as $\mathrm{MAE}=\frac{\sum_{i=1}^{N}\left|y_i-\hat{y}_i\right|}{n}$ . $R^2$ is the percentage of the dependent variable variation that a predictive model explains and defined as $R^2=1-\frac{\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i-\overline{y}_i\right)^2}$.

[8] Random Forest are based on bagging algorithms that aim to control overfitting and reduce variance with independent classifiers. In contrast, Gradient Boosting Machine is based on boosting algorithms which is an approach to reduce bias and variance based on sequential classifiers and increase the complexity of models that suffer from high bias.

[9] There is still a room for improving model's predictive power by scaling input features to a standard range. We can significantly improve the performance measures of each model – in particular, DNN model – by normalizing the standardized features based upon Min-Max Scaler transform as well as Standard Scaler transform, as all input features have the same minimum and maximum values as input for a given algorithm, such as an algorithm that calculates distance measures. However, this will jeopardize the comparisons of such measures between hedonic and machine learning approaches, which is one of the critical goals of this study.

[10] It is worth noting that overfitting problems can often arise from smaller training sets, as the fewer samples for training, the more models can fit the data out. As depicted in Table 4, however, there is not a huge discrepancy in a measure of prediction accuracy between the subsamples and the whole sample. According to the recent machine learning literature, furthermore, overfitting problems can even remain with respectable sample sizes due to other factors, such as number of features relative to the sample size and hyper-parameter optimization (Moghaddam et al., 2020; Vabalas et al., 2019). This implies that it is hard to regularize the relevant sample size in machine learning algorithms, as the sample size does not merely produce optimistically biased results.

[11] We adopt the Mean Decreases Impurity importance measure to measure relative importance of variable to each predictive model Breiman (2001) propose to evaluate the importance of a variable $X_m$ for predicting $Y$ by adding up the weighted impurity decreases $p(t)\Delta i(s_{i},t)$ for all cases $t$ where $X_m$ is used, averaged over all $N_t$ trees in the model. Generaly, splits are defined by a partition of the range $X_m$ of possible values of a single variable $X_m$. The equation for variable importance is defined as follow; $Imp(X_m)=\frac{1}{N_T}\sum_{T}\sum_{t\in T:v(s_t)=X_m} p(t)\Delta i(s_{i},t)$ and where $p(t)$ is the proportion $N_t/N$ of samples reaching $t$ and $v(S_t)$ is the variable used in split $S_t$ .


Random Forest (RF)

Gradient Boosting Machines (GBM)

Support Vector Machines (SVM)

Deep Neural Networks (DNN)

*Note:* Variable importance (VI) represents the statistical significance of variables in each model with respect to its effect on the predictive model. We present variable importance according to the scaled standard importance measure, which rescales the related importance from 1 most critical variable and 0 least critical variable.

Figure 1. Variable importance by machine learning model

model. In addition, it is important to understand the logic of the predictive model not only to verify the accuracy of the model but also to find a way to improve the model by focusing on the important variables. In addition, the variable importance measure enables to select the significant variable and thus has similar performance in much less training time with massive train data set. In Figure 1, variable importance for each predictive mode is depicted. In all predictive models, the assessed value has a higher importance value scaled as 1 followed by building SF as an approximate value of 0.43, 0.39, and 0.57 in RF, GBM, and SVM, respectively. However, we believe that the variable importance from DNN does not provide much useful information about the relative variable importance difference among the variables.

The Random Forest and Support Vector Machine suggest CoStar Rating as the third critical variable but Gradient Boosting Machines suggest land SF as an important variable in order. The GRDP is considered a critical variable at less than 5% importance compared to the assessed value in Random Forest. The investor status as non-local buyer does have an importance value of approximately 12% compared to the assessed value in Support Vector Machines. Interestingly, higher accuracy models in the above analysis, the Random Forest and Gradient Boosting Machines do not consider investor status as the critical value that influence the model prediction.

Also, it is important to note that the critical research variable in this study is investor' status as a local buyer and seller versus non-local buyer and seller. Initially, we establish the local buyer and seller has superior to access local information and network, and thus they will be in an advantageous position when negotiating non-local buyer and seller. Although we find statistically significant empirical evidence in hedonic regression model reported in Table 2, we were not able to find the significant result from variable importance analysis in Figure 1. Only Support Vector Machine suggests the non-local buyer as an important variable predicts the model as the 6th important variable in order. We suspect that the significant result on investor status in hedonic model may be time dependent and focused on specific property types in certain years, such as financial crisis or regional economic shock in a certain time frame, since the significant effect has disappeared in the bagging and boosting method utilizing bootstrapping repeat samplings such as Random Forest and Gradient Boosting Machine. In this sense, we expect from the long-term perspective that this machine learning prediction model may be an alternative tool to make an unbiased estimate for suggestive asking price or initial negotiating price for non-local buyer to overcome possible inferior information accessibility over local investors. In Figure 2, we include a partial dependent plot to show the marginal effect of each variable on the predicted outcome of a machine learning model. A partial dependence plot can show whether the relationship between the target and a feature is linear, monotonic or more complex.

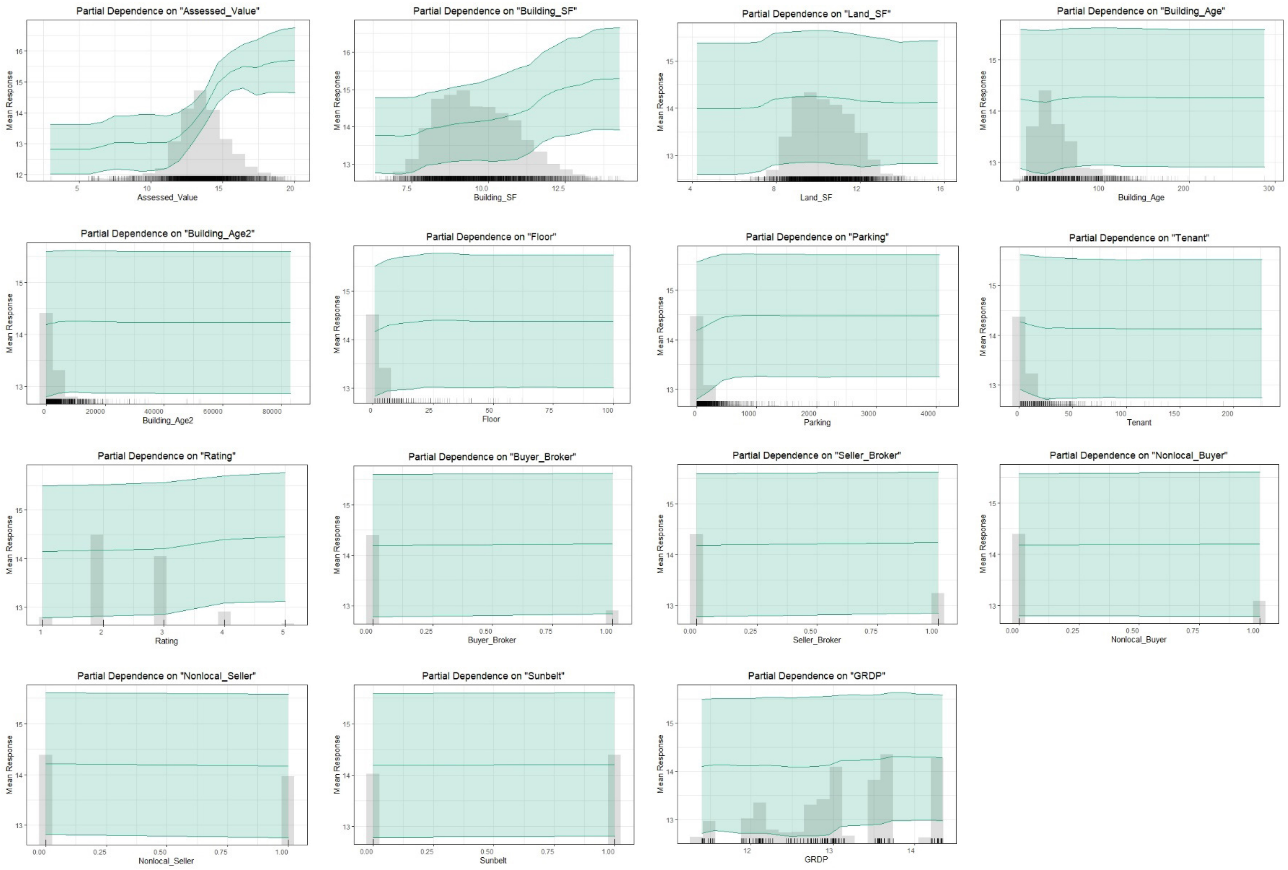
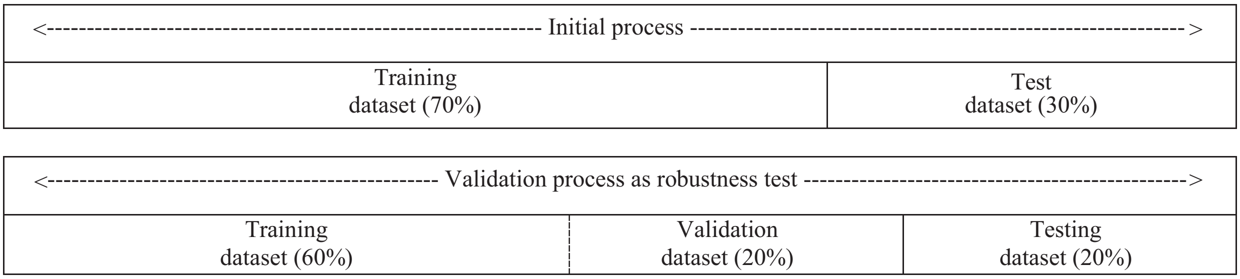


Figure 2. Plots of partial dependence

## 5. Robustness check

In this section, we perform an additional robustness check for our study. At first, we conduct additional analysis for model validation as a robustness check of model accuracy. A validation dataset is a sample of data drawn back from the training model process for the purpose of tuning the parameters of model. The initial analysis described in Table 4 consists of dataset (100%), and we divide the whole dataset into two datasets; training dataset (70%) and test dataset (30%). Thus, the best-fitted model is estimated from the training dataset (70%) predict the test dataset (30%). However, we further examine the validation process with a validation data set. The validation process is useful for estimating the test error related to fitting a predictive model on a training dataset. The validation process randomly divides the available set observations (100%) into a training set (60%) and a validation set (20%), as depicted in the bottom column on Figure 2. The predictive model fits on the training dataset (60%), and the fitted model are supposed to predict the validation dataset (20%) for the purpose of tuning hyperparameters of fitted model from training set. And fitted model after tuning hyperparameters at the validation process predicts the test dataset as a subset of the training dataset. Thus, the clear difference among the training set, validation set, and testing set is that the training set is used for learning to fit the parameters of the classifier while the validation set is used to tune the parameters of the classifier, and test set is used to only to evaluate the performance of the fitted model as depicted in Figure 3.



<-------------------------------------------------------------- Initial process -------------------------------------------------------------->

Training dataset (70%)

Test dataset (30%)

<------------------------------------------------ Validation process as robustness test ------------------------------------------------>

Training dataset (60%)

Validation dataset (20%)

Testing dataset (20%)

Figure 3. A structure of train set, validation set, and test set of ML process

Table 5. Result of RMSE on predictive accuracy models by 10 CMSA in validation and test dataset

| Single CMSA level | RF | | | GBM | | | SVM | | | DNN | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Validation | Test | DIFF | Validation | Test | DIFF | Validation | Test | DIFF | Validation | Test | DIFF |
| Boston | 0.707 | 0.701 | –0.006 | 0.787 | 0.758 | –0.030 | 1.570 | 1.560 | –0.010 | 0.851 | 0.801 | –0.050 |
| Chicago | 0.781 | 0.813 | 0.032 | 0.783 | 0.872 | 0.090 | 1.800 | 1.840 | 0.040 | 1.035 | 1.099 | 0.064 |
| Denver | 0.620 | 0.673 | 0.053 | 0.652 | 0.688 | 0.035 | 1.090 | 1.220 | 0.130 | 0.846 | 0.841 | –0.005 |
| Las Vegas | 0.509 | 0.620 | 0.111 | 0.604 | 0.655 | 0.052 | 0.954 | 0.991 | 0.037 | 0.561 | 0.671 | 0.110 |
| Los Angeles | 0.494 | 0.494 | 0.001 | 0.539 | 0.501 | –0.038 | 0.805 | 0.859 | 0.054 | 0.728 | 0.739 | 0.011 |
| Miami | 0.425 | 0.443 | 0.018 | 0.371 | 0.430 | 0.059 | 1.230 | 0.984 | –0.246 | 0.467 | 0.517 | 0.051 |
| New York | 0.830 | 0.741 | –0.089 | 0.859 | 0.774 | –0.085 | 1.230 | 1.150 | –0.080 | 1.058 | 1.053 | –0.005 |
| San Diego | 0.647 | 0.629 | –0.018 | 1.046 | 0.954 | –0.092 | 0.938 | 1.010 | 0.072 | 0.710 | 0.696 | –0.014 |
| San Francisco | 0.640 | 0.512 | –0.128 | 0.658 | 0.545 | –0.113 | 1.050 | 1.020 | –0.030 | 0.672 | 0.622 | –0.051 |
| Washington D.C | 0.567 | 0.478 | –0.089 | 0.604 | 0.508 | –0.095 | 1.410 | 1.490 | 0.080 | 0.781 | 0.650 | –0.131 |
| 10 CMSA level | | | | | | | | | | | | |
| RMSE | 0.649 | 0.641 | –0.009 | 0.664 | 0.648 | –0.015 | 0.755 | 0.760 | 0.005 | 1.066 | 0.980 | –0.087 |
| MAE | 0.387 | 0.387 | 0.000 | 0.377 | 0.387 | 0.011 | 0.495 | 0.486 | –0.009 | 0.647 | 0.619 | –0.028 |
| MAPE | 2.979 | 2.873 | –0.106 | 2.979 | 2.931 | 0.048 | 3.650 | 3.572 | 0.078 | 4.541 | 4.550 | –0.028 |
| $R^2$ | 0.829 | 0.827 | –0.002 | 0.825 | 0.825 | 0.000 | 0.777 | 0.772 | –0.005 | 0.552 | 0.619 | 0.068 |

*Note:* Root Mean Square Error (RMSE) is the standard deviation of the prediction errors. Since the errors are squared before the errors are averaged, the RMSE gives a relatively high weight to large errors. Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of forecasts, without considering their direction. The MAE measures absolute average difference that are weighted equally in the predictive value. The Mean Absolute Percentage Error (MAPE) is the mean or average of the absolute percentage errors of forecasts. Error is defined as actual or observed value minus the forecasted value and percentage errors are summed without regard to sign to compute MAPE. $R^2$ is the percentage of the dependent variable variation that a predictive model explains and measures the scatter of the sample observation around a predictive model.

Table 6. Result of prediction accuracy in validation and test dataset using 10 high variable importance variables for 10 whole CMSA

| | RF | | | GBM | | | SVM | | | DNN | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 CMSA | Validation | Test | DIFF | Validation | Test | DIFF | Validation | Test | DIFF | Validation | Test | DIFF |
| RMSE | 0.673 | 0.661 | –0.012 | 0.655 | 0.664 | 0.009 | 0.865 | 0.848 | –0.017 | 0.843 | 0.874 | 0.031 |
| MAE | 0.423 | 0.403 | –0.020 | 0.378 | 0.398 | 0.020 | 0.588 | 0.576 | –0.012 | 0.571 | 0.580 | 0.009 |
| MAPE | 3.114 | 2.973 | –0.141 | 2.788 | 2.921 | 0.133 | 4.349 | 4.256 | –0.093 | 4.161 | 4.211 | 0.050 |
| $R^2$ | 0.821 | 0.827 | 0.005 | 0.830 | 0.826 | –0.004 | 0.708 | 0.716 | 0.008 | 0.720 | 0.697 | –0.023 |

Table 5 presents the result of RMSE in validation and test dataset as robustness check with validation process. The validation dataset is used to process for tuning the parameters of the fitted predictive model from the training dataset. Thus, we expect the validation process improves the stability of the predictive model by avoiding the overfitting issue, which implies a possible reduction of the accuracy of the initial model prediction reported in Table 4. We find the difference between validation and test data set is marginal less than 0.01 in random forest and 0.015. In Random Forest in Table 5, as expected, the RMSE measure of the accuracy of test dataset shows 0.641, a lower accuracy than initial process 0.633, and Gradient Boosting Machine shows 0.648, a lower RMSE than 0.633 of RMSE of initial process.

Furthermore, we conducted the validation process with a validation data set using 10 high variable importance derived from Random Forest, Gradient Boosting Machines, Support Vector Machines, and Deep Neural Networks, and the selected variables are depicted in Figure 1. In the ordinary linear regression model generally, many explanatory variables can easily increase the explanation of model accuracy. In this case, model estimation is overestimated and challenging to interpret. To mitigate these concerns, we use ten relatively high-importance variables calculated in Figure 1. Table 6 shows the prediction accuracy results in validation and test dataset using the top 10 variable importance. We find that all models satisfied the difference between validation and test data set is marginally less than 0.05. In Random Forest in Table 6, as expected, the RMSE measure of the accuracy of the test dataset shows 0.661, a lower accuracy than the initial process 0.673, and Gradient Boosting Machine shows 0.664, a lower RMSE than 0.655 of RMSE of the initial process. As a result, we find that it was the same as the results analysed in Tables 4 and 5. The findings from Table 6 provide consistent results compared to our previous results.

## Conclusions

In this study, we examine the machine learning method to measure the prediction accuracy for commercial real estate transaction prices. Previous researches attempt to discuss a machine learning methodology and its application on the residential market (Mullainathan & Spiess, 2017; Ho et al., 2021). We contribute to this literature by providing an accuracy measure based on the machine learning method in the commercial real estate market. Utilizing total 19,640 CoStar office transaction data covering for national level and 10-CMSA level over 14 years of data length, we attempt to analyze machine learning methods; Random Forest and Support Vector Machine to estimate the best prediction model for commercial real estate transaction price. We also contribute to the existing literature by providing empirical evidence of investor status as non-local buyer and non-local seller in the context of machine learning framework. The result from machine learning with superior accuracy over the hedonic regression model will be an unbiased estimate for non-local investor in the commercial real estate market. While we not only literally provide an accuracy of each prediction model, we also provide a relatively marginal influence to each prediction model to verify the relative importance of each variable.

The major finding of our study is that there is a significant difference between the conventional hedonic approach and machine learning methods: Random Forest, Gradient Boosting Machine, Support Vector Machine, and Deep Neural Networks. Consistent with a recent study in the residential market by Ho et al. (2020), we also find the Random Forest and Gradient Boosting Machine performed best in office property transaction prices at both national and CMSA levels.

In the analysis of hedonic regression model, the finding suggests that investor status is statistically significant to the transaction price. Consistent with previous literature (Liu et al., 2015; Ling et al., 2018), non-local buyers systematically pay a premium when they transact office property, and non-local seller systematically sold an office property with a discount when they transact. In hedonic regression analysis, we also find statistically significant property specific variables such as assessed value, building size, building age, floor, number of parking, number of tenants, and CoStar rating. Thus, we can confirm that non-local buyer pays a premium on office property transaction compared to that local buyer when they purchase, and non-local seller transacted an office property with discount compared to local seller when they sell an office property in the hedonic regression analysis. However, this investor status effect on transaction price has disappeared or at least dampened in optimized machine learning estimate. We find that Random Forest and Gradient Boosting Model as the best predictive model but none of these models suggest a variable of investor status as at least the 10th important variable. We attribute an inconsistent outcome to the characteristics of each machine learning method. The Random Forest and Gradient Boosting Machine are both based on process of sampling from training data set and then applies it to test dataset to adjust the prediction model.

In hedonic model, there might be an increase in non-local buyer transactions during a regional economic downturn or time period of increase transactions. Thus, the transaction associated non-local investor as a dataset can be treated as time-specific and regional-specific isolated dataset and treated as dummy variables that distinguish its effect from other variables. On the other hand, the observation randomly drawn from the training data set may contain general characteristics to accurately explain its test dataset. The possible time-specific and regional-specific investor's effect might be generalized in the optimization process. We believe this logic is also applicable similarly to the validation process. The search process to optimize model is based on bagging algorithm and boosting algorithms. If the sample observations of the predictive model represent a general characteristic of each model, then the effect transactions associated with non-local investors may decay, and the predicted value can be an unbiased estimate for non-local investors in the future transaction.

Thus, if non-local buyers and sellers are sufficiently informed on the possible transaction price estimated by machine learning method, then a possible premium paid by non-local buyer and discount charged by non-local seller will be dampened or at least reduced. However, our study also has a limitation in practice. Although we attempt to optimize the prediction model process with a set choice and value range of variable size and number, we were not able to provide the best or most practical optimization condition persistent for future trials. Also, we follow to make a partition of the training data set and test data set, but it is still not also standardized and still needs to be examined. Also, in a practical sense, there might be a trade-off between a direct interpretation of the magnitude of causal effect by interpreting the coefficient in a hedonic framework and better prediction accuracy in the machine learning method with an ambiguous interpretation of direct causal effect on the dependent variable.

To conclude, due to increased information in the commercial real estate market, we were able to attempt to analyze a new analytical machine learning method to estimate commercial real estate market transaction prices. Since the increased amount of accessible qualitative commercial data, we were able to compare the new machine learning method and the traditional hedonic approach. We believe that our study will be a groundwork for further discussion on methodological innovation to the valuation of the commercial real estate market with ongoing innovation in commercial real estate datasets and technological innovation.

## Acknowledgements

The authors are grateful for constructive comments raised by reviewers as well as reviewers who attended 2021 Asian Real Estate Society Annual Conference held in Singapore.

## Funding

This work was not supported by any financial support.

## Author contributions

Jin and Jung conceived the study and were responsible for the design and development of the data analysis. Jung, Jin and Kim were responsible for data collection and analysis. Kim and Jin were responsible for data interpretation.

## Disclosure statement

Authors do not have any competing financial, professional, or personal interests from other parties.

## References

Bishop, C. M. (2006). Information science and statistics. In *Pattern recognition and machine learning*. Springer.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge. https://doi.org/10.1201/9781315139470

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information*, *7*(5), 168. https://doi.org/10.3390/ijgi7050168

Colwell, P., Munneke, H., & Trefzger, J. (1998) Chicago's office market: price indices, location and time. *Real Estate Economics*, *26*(1), 83–106. https://doi.org/10.1111/1540-6229.00739

Conway, J. J. E. (2018). *Artificial intelligence and machine learning: current applications in real estate*. https://dspace.mit.edu/handle/1721.1/120609

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cowden, C., Fabozzi, F. J., & Nazemi, A. (2019). Default prediction of commercial real estate properties using machine learning techniques. *The Journal of Portfolio Management*, *45*(7), 55–67. https://doi.org/10.3905/jpm.2019.1.104

Egan, M. (2019, February 17). How elite investors use artificial intelligence and machine learning to gain an edge. *CNN Business*. https://edition.cnn.com/2019/02/17/investing/artificial-intelligence-investors-machine-learning

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gallimore, P., & Wolverton, M. (2000). The objective in valuation: a study of the influence of client feedback. *Journal of Property Research*, *17*(1), 47–57. https://doi.org/10.1080/095999100368010

Geltner, D., MacGregor, B. D., & Schwann, G. M. (2003). Appraisal smoothing and price discovery in real estate markets. *Urban Studies*, *40*(5–6), 1047–1064. https://doi.org/10.1080/0042098032000074317

Gupta, R., Marfatia, H. A., Pierdzioch, C., & Salisu, A. A. (2022). Machine learning predictions of housing market synchronization across US states: the role of uncertainty. *The Journal of Real Estate Finance and Economics*, *64*, 523–545. https://doi.org/10.1007/s11146-020-09813-1

Han, L., & Hong, S. H. (2016). Understanding in-house transactions in the real estate brokerage industry. *The RAND Journal of Economics*, *47*(4), 1057–1086. https://doi.org/10.1111/1756-2171.12163

Hansz, J. A., & Diaz III, J. (2001). Valuation bias in commercial appraisal: a transaction price feedback experiment. *Real Estate Economics*, *29*(4), 553–565. https://doi.org/10.1111/1080-8620.00022

Hill, R. J. (2013). Hedonic price indexes for residential housing: a survey, evaluation and taxonomy. *Journal of Economic Surveys*, *27*(5), 879–914.

Ho, W., Tang, B., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107–116. https://doi.org/10.1142/S0218488598000094

International Association of Assessing Officers. (2013). *Standard on mass appraisal of real property*. https://www.iaao.org/media/standards/StandardOnMassAppraisal.pdf

Izmailov, R., Vapnik, V., & Vashist, A. (2013). Multi-dimensional splines with infinite number of knots as SVM kernels. In *The 2013 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–7). IEEE. https://doi.org/10.1109/IJCNN.2013.6706860

Kandlbinder, K. (2018). *The role of information in real estate markets* [Doctoral dissertation, Universität Regensburg]. https://epub.uni-regensburg.de/37492/1/00%20Dissertation_Pflichtexemplare.pdf

Kok, N., Koponen, E. L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *The Journal of Portfolio Management*, *43*(6), 202–211. https://doi.org/10.3905/jpm.2017.43.6.202

Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-801460-8.00013-6

Lam, K. C., Yu, C. Y., & Lam, C. K. (2009). Support vector machine and entropy-based decision support system for property valuation. *Journal of Property Research*, *26*(3), 213–233. https://doi.org/10.1080/09599911003669674

Ling, D. C., Naranjo, A., & Scheick, B. (2018). Geographic portfolio allocations, property selection and performance attribution in public and private real estate markets. *Real Estate Economics*, *46*(2), 404–448. https://doi.org/10.1111/1540-6229.12184

Liu, Y., Gallimore, P., & Wiley, J. A. (2015). Non-local office investors: anchored by their markets and impaired by their distance. *The Journal of Real Estate Finance and Economics*, *50*(1), 129–149. https://doi.org/10.1007/s11146-013-9446-8

McMillen, D. P., & Redfearn, C. L. (2010). Estimation and hypothesis testing for nonparametric hedonic house price functions. *Journal of Regional Science*, *50*(3), 712–733. https://doi.org/10.1111/j.1467-9787.2010.00664.x

Meese, R. A., & Wallace, N. E. (1991). Nonparametric estimation of dynamic hedonic price models and the construction of residential housing price indices. *Real Estate Economics*, *19*(3), 308–332. https://doi.org/10.1111/1540-6229.00555

Moghaddam, D. D., Rahmati, O., Panahi, M., Tiefenbacher, J., Darabi, H., Haghizadeh, A., Haghighi, A. T., Nalivang, O. A., & Bui, D. T. (2020). The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers. *Catena*, *187*, 104421. https://doi.org/10.1016/j.catena.2019.104421

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Nghiep, N., & Al, C. (2001). Predicting housing value: a comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research*, *22*(3), 313–336. https://doi.org/10.1080/10835547.2001.12091068

Parmeter, C. F., Henderson, D. J., & Kumbhakar, S. C. (2007). Nonparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, *22*(3), 695–699. https://doi.org/10.1002/jae.929

Pérez-Rave, J. I., Correa-Morales, J. C., & González-Echavarría, F. (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research*, *36*(1), 59–96. https://doi.org/10.1080/09599916.2019.1587489

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, *31*(2), 147–164. https://doi.org/10.1080/10835547.2009.12091245

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301. https://doi.org/10.1002/widm.1301

Rafiei, M. H., & Adeli, H. (2016). A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management*, *142*(2), 04015066. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001047

Rice, L., Wong, E., & Kolter, Z. (2020). Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning* (pp. 8093–8104). PMLR.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. https://doi.org/10.1007/BF00116037

Schapire, R., & Freund, Y. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Second European Conference on Computational Learning Theory* (pp. 23–37). Springer. https://doi.org/10.1007/3-540-59119-2_166

Shen, L., & Ross, S. (2021). Information value of property description: a machine learning approach. *Journal of Urban Economics*, *121*, 103299. https://doi.org/10.1016/j.jue.2020.103299

Shi, L., & Tapia, C. (2016). The disciplining effect of concern for referrals: evidence from real estate agents. *Real Estate Economics*, *44*(2), 411–461. https://doi.org/10.1111/1540-6229.12102

Simlai, P. E. (2021). Predicting owner-occupied housing values using machine learning: an empirical investigation of California census tracts data. *Journal of Property Research*, *38*(4), 305–336. https://doi.org/10.1080/09599916.2021.1890187

Sun, X., Ren, X., Ma, S., & Wang, H. (2017). meProp: sparsified back propagation for accelerated deep learning with reduced overfitting. In *International Conference on Machine Learning* (pp. 3299–3308). PMLR.

Telgarsky, M. (2013). Margins, shrinkage, and boosting. In *International Conference on Machine Learning* (pp. 307–315). PMLR.

Turnbull, G. K., & Sirmans, C. F. (1993). Information, search, and house prices. *Regional Science and Urban Economics*, *23*(4), 545–557. https://doi.org/10.1016/0166-0462(93)90046-H

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS ONE*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Wong, S. K., Yiu, C. Y., & Chau, K. W. (2012). Liquidity and information asymmetry in the real estate market. *The Journal of Real Estate Finance and Economics*, *45*(1), 49–62. https://doi.org/10.1007/s11146-011-9326-z

Xu, H., & Gade, A. (2017). Smart real estate assessments using structured deep neural networks. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation* (pp. 1–7). IEEE. https://doi.org/10.1109/UIC-ATC.2017.8397560

Yao, Y., Zhang, J., Qian, C., Wang, Y., Ren, S., Yuan, Z., & Guan, Q. (2021). Delineating urban job-housing patterns at a parcel scale with street view imagery. *International Journal of Geographical Information Science*, *35*(10), 1927–1950. https://doi.org/10.1080/13658816.2021.1895170

Yilmazer, S., & Kocaman, S. (2020). A mass appraisal assessment study using machine learning based on multiple regression and random forest. *Land Use Policy*, *99*, 104889. https://doi.org/10.1016/j.landusepol.2020.104889

You, Q., Pang, R., Cao, L., & Luo, J. (2017). Image-based appraisal of real estate properties. *IEEE Transactions on Multimedia*, *19*(12), 2751–2759. https://doi.org/10.1109/TMM.2017.2710804

Yu, L., Jiao, C., Xin, H., Wang, Y., & Wang, K. (2018). Prediction on housing price based on deep learning. *International Journal of Computer and Information Engineering*, *12*(2), 90–99.

Zhou, X., Gibler, K., & Zahirovic-Herbert, V. (2015). Asymmetric buyer information influence on price in a homogeneous housing market. *Urban Studies*, *52*(5), 891–905. https://doi.org/10.1177/0042098014529464