CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS

INFOTEC

BIBLIOTECA INFOTEC
VISTO BUENO DE TRABAJO TERMINAL

Maestría en Ciencia de Datos e Información
(MCDI)

Ciudad de México, a 12 de noviembre de 2024

UNIDAD DE POSGRADOS
PRESENTE

Por medio de la presente se hace constar que el trabajo de titulación:

"Machine Learning Applications to Property Price Indexes: A Market Comparative Analysis"

Desarrollado por el alumno: **Juan Carlos Antonio Téllez Velasco,** bajo la asesoría del **Dr. Mario Graff Guerrero,** cumple con el formato de Biblioteca, así mismo, se ha verificado la correcta citación para la prevención del plagio; por lo cual, se expide la presente autorización para entrega en digital del proyecto terminal al que se ha hecho mención. Se hace constar que la aluna no adeuda materiales de la biblioteca de INFOTEC.

No omito mencionar, que se deberá anexar la presente autorización al inicio de la versión digital del trabajo referido, con el fin de amparar la misma.

Sin más por el momento, aprovecho la ocasión para enviar un cordial saludo.

Mtro. Carlos Josué Lavandeira Portillo
Director Adjunto de Innovación y Conocimiento

CJLP/dqor

C.c.p. Felipe Alfonso Delgado Castillo.- Gerente de Capital Humano.- Para su conocimiento.
Juan Carlos Antonio Téllez Velasco.- Alumno de la Maestría en Ciencia de Datos e Información.- Para su conocimiento.

MCDI
MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO
GERENCIA DE CAPITAL HUMANO
POSGRADOS

“Machine Learning Applications to Property Price Indexes: A Market Comparative Analysis”

TESIS
Que para obtener el grado de MAESTRO EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

Juan Carlos Antonio Téllez Velasco

Asesor:

Dr. Mario Graff Guerrero

Ciudad de México, Mayo, 2024.

CONAHCYT

INFOTEC POSGRADOS

Index

Figures Index

Table Index

Acronyms

ML	Machine Learning
IA	Artificial Intelligence
MSCI	Morgan Stanley Capital International
RCA	Real Capital Analytics
OLS	Ordinary Least Squares
RMSE	Root Mean Squared Error
MSE	Mean Squared Error
SD	Standard Deviation
CP	Chained Paasche Index
SVR	Support Vector Regression
RF	Random Forest
DT	Decision Tree
XGB	Extreme Gradient Boosting

Glossary

Algorithm: A set of instructions or logical rules designed to solve a problem or perform a specific task (Investopedia, 2023a).

Big Data: A collection of data so large and complex that it is difficult to process using conventional data analysis tools (Liebowitz, 2013).

Index: A statistical measure or indicator that represents the performance or value of a specific sector or market (Investopedia, 2023c).

Machine Learning: Algorithms that enable computer systems to learn and make predictions or decisions without being explicitly programmed (DeepConverse, 2024).

R squared (R^2): Statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable (Investopedia, 2024b).

Real Assets: Physical or tangible assets that have intrinsic value due to their substance and properties (Investopedia, 2024a).

Real Estate: Real property that includes land and anything permanently attached to it or built on it, whether natural or man-made (Investopedia, 2024c).

Returns: The money made or lost on an investment over some period of time (Investopedia, 2023b)

RMSE (Root Mean Squared Error): A measure of the differences between values that are predicted by a model and values that are actually observed (Gupta et al., 2022).

Test Set: A subset of data used to evaluate the performance of a machine learning model after it has been trained (Isholafa, 2024).

Training Set: A subset of data used to train a machine learning model (Isholafa, 2024)

Volatility: Price variation of an asset over time (Jurczenko, 2017).

Abstract

This research studies the application of machine learning algorithms in estimating property price indexes. It expands on the work of Calainho et al. (2022) and evaluates the effectiveness of these algorithms (Extreme Gradient Boosting, Support Vector Machine, Random Forest and Decision Tree) in the cities of Atlanta, Chicago, Los Angeles, Miami, and Phoenix, from 2000 to 2024. The analysis is based on over 200,000 property transactions and makes over 600 estimates to assess the accuracy of these algorithms across five different cities. This research analyzes various cities to understand their impact on the algorithm's accuracy and contributes to the discussion on methodologies for estimating real estate indexes, essential for the industry.

Resumen

Este trabajo estudia la aplicación de algoritmos de aprendizaje automático en la estimación de índices de precios de propiedades. Amplía el trabajo de Calainho et al. (2022) y evalúa la efectividad de diversos algoritmos de aprendizaje automático (*Extreme Gradient Boosting, Support Vector Machine, Random Forest* y *Decision Tree*) en las ciudades de Atlanta, Chicago, Los Ángeles, Miami y Phoenix, desde 2000 hasta 2024. El análisis se basa en más de 200,000 transacciones de propiedades y realiza más de 600 estimaciones para evaluar la precisión de estos algoritmos en cinco ciudades. El trabajo analiza diversas ciudades para comprender su impacto en la precisión de los algoritmos y contribuye a la discusión sobre metodologías para estimar índices de bienes raíces, esenciales para la industria.

Introduction

Real estate serves as a major asset class in the investment industry and plays a crucial role in the global economy. It provides stability and potential for capital and income return as an investment vehicle. Real estate is essential for wealth creation and preservation not only for individual investors but also institutional investors, governments, and corporations. Its impact on economic growth, financial stability, and infrastructure underlines its importance.

According to Fenwick (2013), property price indexes serve multiple purposes. They act as indicators of economic growth and aid in monetary policy. Indexes estimate the value of housing and serve as stability indicators to measure risk exposure. Furthermore, property price indexes are used as deflators in national accounts and as inputs for individual citizens' decision-making on whether to buy or sell property. Additionally, property price indexes play a crucial role in the consumer price index, used for indexation purposes. They are instrumental in making urban and international comparisons. As a result, the multiple uses of property price indexes highlight their importance in policy, and economic and financial analysis, making them essential to understand the real estate market.

This research builds on the work of Calainho et al. (2022), which introduced a methodology for creating property price indexes using machine learning algorithms, focusing on real estate transactions in New York from 2000 to 2019. This research expands on their work and evaluates the effectiveness of machine learning algorithms in the cities of Atlanta, Chicago, Los Angeles, Miami, and Phoenix, from 2000 to 2024. This study aims to understand the robustness of machine learning in estimating property price indexes across different cities to encourage the use of the algorithms. The research employs a dataset comprising more than 200,000 property transactions. This dataset serves as a robust basis for evaluating the effectiveness of these models.

Additionally, this research paper builds on the authors' work by using similar machine learning algorithms, such as Extreme Gradient Boosting and Support Vector Machine, and adds Random Forest and Decision Tree, but excludes Artificial Neural Networks and Cubist algorithms from the analysis. It adopts a single imputation approach and does not use expanding or rolling window calculation methods. Additionally, it does not provide sample size analysis. The main contribution of this research, compared to that of the authors, is the market comparison analysis to study the impact of location-specific factors on model accuracy.

The adoption of artificial intelligence technologies is transforming various sectors. However, the real estate industry has been slower to adopt it, even though it increasingly uses analytics to extract insights from datasets. The use of machine learning algorithms in estimating property price indexes beyond academia could provide a helpful push to move forward. This research aims to start a discussion on considering different methodologies for product development.

The document is structured as follows: It starts with a review of existing literature, followed by an overview of the dataset used. The methodology section introduces the concept of chained indexes and the machine learning models used. The results section discusses index estimates, and market comparison analysis to assess the models' performance. The conclusion summarizes the limitations, main findings, and their significance for research on property price indexes.

Chapter 1
Literature Review

Literature Review

This section provides an overview of the related literature to this research. Hill (2011) considers the median index as the simplest property index, measuring the change in median dwelling prices over time. Median indexes are popular because they require less data and are easy to calculate. However, their main disadvantage is that they mix price and quality changes, leading to imprecise estimates of price changes. This problem arises because the median quality can vary between periods, causing the index to reflect shifts that do not accurately represent the overall market, as seen when comparing sales from different areas with varying property values over consecutive years. Furthermore, Costa & Cazassa (2017) suggest that while median-price indexes are important, they may exhibit higher volatility during market peaks and troughs, especially in areas with diverse property mixes. The authors recommend considering this bias when interpreting median-price indexes and suggest using hedonic-based techniques or stratification by location and property characteristics as solutions.

Property Price Indexes

Bailey et al. (1963) introduced the repeat-sales method, which tracks changes in property prices over time using data from properties sold at least twice. This method assumes that price changes between sales reflect general market trends rather than specific property improvements, ensuring comparability. However, Hill (2011) notes that this approach can lead to selection bias, since price movements of these properties might not represent the overall market, leading to biases in the repeat-sales indexes during economic fluctuations. Melser (2023) proposed the Characteristics Repeat Sales (CRS) method to overcome this limitation. When applied to data from Florida between 2002 and 2020, the CRS index demonstrated less volatility compared to the standard repeat-sales index, suggesting that the CRS method offers a more accurate reflection of price changes.

The hedonic approach assesses the price of a product based on its characteristics, making it suitable for creating quality-adjusted price indexes in markets like housing. Real estate attributes can be divided into physical features, such as the number of bedrooms land area and locational features (Owusu-Ansah, 2011). Coulson (2008) identifies three uses of hedonic methods: constructing quality-adjusted property price indexes, providing automated property valuations, and analyzing property price change due to amenities and disamenities. Hill (2011) states that the hedonic method serves two purposes: adjusting observed prices for quality differences to create price index and estimating consumers' willingness to pay for specific product features. The hedonic approach often faces challenges with omitted variables affecting estimations. Rosen (1974) points that observed prices and quantities are outcomes of supply and demand equilibrium, making it difficult to isolate and determine causal relationships between variables.

Machine Learning & Real Estate

Furthermore, researchers have widely used artificial intelligence and machine learning approaches in real estate studies. Kok et al. (2017) suggests using machine learning to automated valuation models. They show that these models can be more accurate and cost-effective than traditional appraisals methods. Similarly, Tay and Ho (1992) used artificial neural networks and discuss their application to valuing residential apartments, comparing the performance of a back-propagation artificial neural network model against the traditional regression model. The authors concluded that the artificial neural network model provides an alternative, though it remains a black-box approach.

Evans et al. (1995) examine the application of the same method on 288 home sales in Colorado. The study finds that artificial neural networks are not necessarily superior for appraisal analysis, noting significant issues with inconsistent results and computing complexity issues.

Nghiep & Al (2001) compare the predictive performance of artificial neural networks and regression analysis for single-family housing sales, with variations in data

sample size. The study finds that neural networks outperform regression analysis with a moderate to large data sample size.

Peterson & Flanagan (2009) analyze a large sample of residential properties from 1999-2005 and find that neural networks outperform linear hedonic pricing models in terms of lower dollar pricing errors, greater out-of-sample pricing precision, and better extrapolation from volatile pricing environments. They highlight that multiple-layered neural networks can model complex nonlinearities and are well-suited to hedonic models with large numbers of dummy variables.

Real estate research has seen increased use of machine learning. Choy and Ho (2023) showed that algorithms like Extra Trees, k-Nearest Neighbors and Random Forest can better predict property prices than traditional statistical techniques, showing its ability to give more accurate price predictions in real estate. Viriato (2019) looked at how machine learning could change real estate investment, noting that over 100 real estate technology companies use these methods to improve insights, productivity, and accuracy. Additionally, Baldominos et al. (2018) created a machine learning tool to find real estate opportunities in real time, especially properties listed below market price, which can help investors in the housing market. These studies show machine learning's various uses and potential in real estate, from price prediction to finding investment opportunities.

Machine Learning & Property Price Indexes

As it was mentioned before, this research builds on the work of Calainho et al. (2022), which introduced a methodology for creating property price indexes using machine learning algorithms, focusing on real estate transactions in New York from 2000 to 2019. The authors present an approach that uses out-of-time individual transaction predictions to build price indexes using different machine learning algorithms.

The research highlights the need to address estimation bias in index construction. The authors underline the significance of hyperparameter tunning in minimizing bias and avoiding overfitting. Their stress tests reveals that ordinary least square (OLS)

generate more stable indexes with limited training data compared to non-linear machine learning algorithms, as they depend less on the number of observations. OLS also shows lower index volatility and smaller variations in the loss function across different data availability levels than machine learning algorithms. The study finds that machine learning algorithms are more dependent on sample size.

The analysis of the optimal window size for the rolling window approach shows significant variation across different algorithms, with window sizes ranging from 2 to 8 years. The authors found that the single-year window approach has a higher RMSE compared to the rolling window and expanding window approaches. Overall, the authors conclude that machine learning algorithms tend to produce better results than OLS when more observations are available, even if it means adding more dimensions. In contrast, OLS performs better in datasets with fewer observations or characteristics.

Chapter 2
Data

Data

The data consists of property transactions in Atlanta, Chicago, Los Angeles, Miami, and Phoenix, over the period 2000–2024. The data comes from MSCI, Inc.[1] For each property transaction, the data contains transaction price in United States dollars (USD), the area in squared feet, region, property type and building year. Having access to transaction data is a significant strength in this research.

Kolbe et al. (2021) considered using real estate platform listing data as a substitute for transaction data in hedonic regression applications. However, they found that asking prices are generally higher than sale prices. Additionally, using listing data resulted in significant discrepancies in willingness-to-pay estimates and proved less reliable in predicting property prices due to upward bias and high error variance.

In the preprocessing of the data, several steps were taken to ensure the quality of the data for analysis. First, any rows with missing values were removed to maintain the integrity of the dataset. Next, a categorical variable for the building period was created by dividing the construction year data into quartiles. Dummy variables were created for property types and regions. The dataset was filtered based on a ratio calculated as the logarithm of the price divided by the logarithm of the area. Only data points contained in the 1st and 99th percentiles of this ratio were retained, removing potential outliers. Finally, a logarithmic transformation was applied to the transaction price.

Table 1 presents a yearly summary of real estate data for Los Angeles from 2000 to 2024. It includes average and standard deviation of price and area, and the number of observations each year. The data shows that both the average price and area changes over time. There is a noticeable [redacted] increase in the average price from 2000 to 2024. Similarly, the standard deviation of the price also varies significantly,

[1] Certain information ©2024 MSCI ESG Research LLC. Reproduced by permission.

indicating high volatility in prices. The average area has been steadily decreasing. Also, the number of observations peaks in [redacted] at [redacted]

Table 1 also categorizes the real estate data of Los Angeles into regions, property types, and building periods. The number of observations indicates that Los Angeles is the region with the most transactions at [redacted] while Ventura Co with [redacted] Regarding property types, [redacted] lead with [redacted] of the observations, whereas [redacted] have the fewest. When breaking down by building period, the majority of properties transacted were constructed between [redacted] amounting to [redacted] of the observations. This table gives a clear indication of market trends in real estate within various categories over time.

We can infer the dynamics of the Los Angeles real estate market over a 24-year period. The data shows an increase in the average price of properties despite volatility in the market. The peak in observations in [redacted] suggests a busy market that has since seen a decrease in activity by 2024. Moreover, the categorical variables provide a detailed look at market segmentation, showing that most transactions took place in the Los Angeles region, with Apartments being the most popular property type and the majority of properties having been built in the last decades.

Tables 2 to 5 present data from Atlanta, Chicago, Miami, and Phoenix, sourced from MSCI, Inc. The following pages present this data. Commentary has been omitted to prevent redundancy, aligning with the structure used for Los Angeles. These markets exhibit comparable macroeconomic trends, yet distinct nuances arise from differences in population size and consumer and productive amenities. This research also omitted an exploratory analysis and opted for concise tabular data presentation to streamline information.

Los Angeles Data

Table 1. Los Angeles Market Summary Statistics[2]

Year	Avg. Price	SD. Price	Avg. Area	SD. Area	Obs.
2000	$	$			
2001	$	$			
2002	$	$			
2003	$	$			
2004	$	$			
2005	$	$			
2006	$	$			
2007	$	$			
2008	$	$			
2009	$	$			
2010	$	$			
2011	$	$			
2012	$	$			
2013	$	$			
2014	$	$			
2015	$	$			
2016	$	$			
2017	$	$			
2018	$	$			
2019	$	$			
2020	$	$			
2021	$	$			
2022	$	$			
2023	$	$			
2024	$	$			

Los Angeles Categorical Variables

Regions	Obs.	Property Type	Obs.	Building Period	Obs.
Inland Empire		Apartment		1 - <1962	
Los Angeles		Dev Site		2 - 1962-1978	
Orange Co		Hotel		3 - 1978 to 1990	
Ventura Co		Industrial		4 – 1990>	
		Office			
		Other			
		Retail			
		Seniors Housing & Care			

Total =

[2] Avg. Price = Average Price in US Dollars; SD. Price = Price Standard Deviation in US Dollars; Avg. Area = Average Area in Square Feet; SD. Area = Area Standard Deviation in Square Feet; Obs. = Number of Observations

Atlanta Data

Table 2. Atlanta Market Summary Statistics[3]

Year	Avg. Price	[redacted]	Avg. Area	SD. Area	Obs.
2000	$				
2001	$				
2002	$				
2003	$				
2004	$				
2005	$				
2006	$				
2007	$				
2008	$				
2009	$				
2010	$				
2011	$				
2012	$				
2013	$				
2014	$				
2015	$				
2016	$				
2017	$				
2018	$				
2019	$				
2020	$				
2021	$				
2022	$				
2023	$				
2024	$				

Atlanta Categorical Variables

Regions	Obs.	Property Type	Obs.	Building Period	Obs.
Atlanta		Apartment		1 - <1980	
		Dev Site		2 – 1980-1994	
		Hotel		3 – 1994-2003	
		Industrial		4 – 2003>	
		Office			
		Other			
		Retail			
		Seniors Housing & Care			
Total =					

[3] Avg. Price = Average Price in US Dollars; SD. Price = Price Standard Deviation in US Dollars; Avg. Area = Average Area in Square Feet; SD. Area = Area Standard Deviation in Square Feet; Obs. = Number of Observations

Chicago Data

Table 3. Chicago Market Summary Statistics[4]

Year	Avg. Price	[redacted]	Avg. Area	SD. Area	Obs.
2000	$	$			
2001	$	$			
2002	$	$			
2003	$				
2004	$				
2005	$				
2006	$				
2007	$				
2008	$				
2009	$				
2010	$				
2011	$				
2012	$				
2013	$				
2014	$				
2015	$				
2016	$				
2017	$				
2018	$				
2019	$				
2020	$				
2021	$				
2022	$				
2023	$				
2024	$				

Chicago Categorical Variables

Regions	Obs.	Property Type	Obs.	Building Period	Obs.
Chicago		Apartment		1 - <1960	
Kankakee		Dev Site		2 - 1960-1984	
		Hotel		3 – 1984-2001	
		Industrial		4 – 2001>	
		Office			
		Other			
		Retail			
		Seniors Housing & Care			
Total =					

[4] Avg. Price = Average Price in US Dollars; SD. Price = Price Standard Deviation in US Dollars; Avg. Area = Average Area in Square Feet; SD. Area = Area Standard Deviation in Square Feet; Obs. = Number of Observations

Miami Data

Table 4. Miami Market Summary Statistics[5]

Year	Avg. Price	SD. Price	Avg. Area	SD. Area	Obs.
2000	$				
2001	$				
2002	$				
2003	$				
2004	$				
2005	$				
2006	$				
2007	$				
2008	$				
2009	$				
2010	$				
2011	$				
2012	$				
2013	$				
2014	$				
2015	$				
2016	$				
2017	$				
2018	$				
2019	$				
2020	$				
2021	$				
2022	$				
2023	$				
2024	$	$			

Miami Categorical Variables

Regions	Obs.	Property Type	Obs.	Building Period	Obs.
Broward		Apartment		1 - <1968	
Miami/Dade Co		Dev Site		2 - 1968-1985	
Palm Beach Co		Hotel		3 – 1985-2001	
		Industrial		4 – 2001>	
		Office			
		Other			
		Retail			
		Seniors Housing & Care			
Total =					

[5] Avg. Price = Average Price in US Dollars; SD. Price = Price Standard Deviation in US Dollars; Avg. Area = Average Area in Square Feet; SD. Area = Area Standard Deviation in Square Feet; Obs. = Number of Observations

Phoenix Data

Table 5. Phoenix Market Summary Statistics[6]

Year	Avg. Price	SD. Price	Avg. Area	SD. Area	Obs.
2000					
2001					
2002					
2003					
2004					
2005					
2006					
2007					
2008					
2009					
2010					
2011					
2012					
2013					
2014					
2015					
2016					
2017					
2018					
2019					
2020					
2021					
2022					
2023					
2024					

Phoenix Categorical Variables

Regions	Obs.	Property Type	Obs.	Building Period	Obs.
Phoenix		Apartment		1 - <1982	
		Dev Site		2 – 1982-1995	
		Hotel		3 – 1995-2005	
		Industrial		4 – 2005>	
		Office			
		Other			
		Retail			
		Seniors Housing & Care			

Total =

[6] Avg. Price = Average Price in US Dollars; SD. Price = Price Standard Deviation in US Dollars; Avg. Area = Average Area in Square Feet; SD. Area = Area Standard Deviation in Square Feet; Obs. = Number of Observations

Chapter 3
Methodology

Methodology

This research uses a chained index approach, and both linear and non-linear models to estimate the value of real estate properties over different time periods. The linear model, estimated using OLS, is used as a benchmark. For non-linear modeling, the paper applies four machine learning algorithms, including Support Vector Regression, Decision Tree, Random Forest and Extreme Gradient Boosting.

Chained Paasche Index

The chained hedonic method uses a sequence of property prices estimations over multiple years. The methodology controls for changes in sample size in each period by specifying a representative property (the average property i in period t). The model does not use time fixed effects, as the prices are estimated in each period, making time a constant. The equation is estimated using log sale prices (P_t) and characteristics (X_t). The vectors β'_t and ε_t correspond to the estimates of the parameter vector and the error vector, respectively. Matrices are denoted in bold throughout the text. We use base period (t) for training, and then predict log prices of the following period (s):

TRAINING

$$\boldsymbol{P_t} = \mathbf{X_t}\beta'_t + \varepsilon_t$$

PREDICTION

$$\boldsymbol{P'_s} = \mathbf{X_s}\beta'_t + \varepsilon_s$$

This research utilizes a single imputation approach, employing actual transaction prices along with estimated prices to calculate the price difference. While Balk et al. (2013) advocate for double imputation as the optimal method for estimating indexes—due to its ability to mitigate biases from omitted variables by offsetting errors from different imputations—we argue that our method is easier to implement and aligns better with practical machine learning applications.

The average difference between the estimated log sale prices and the actual prices during the period, where N_s represents the total number of properties in the dataset and n_s the number of properties in the current period, provides an estimate of the log price change:

$$\delta'_s = \frac{1}{n_s} \sum_{i \in N_s} (P_{is} - P'_{is})$$

Finally, the Chained Paasche Index (CP), which we set at 100 for its initial value, for the current period (s) is calculated by multiplying the index from the base period (t) by the exponential of the estimated log price difference:

$$\mathrm{CP_s} = \mathrm{CP_t} \times \mathrm{e}^{\delta'_s}$$

Machine Learning

Machine learning aims to train algorithms to recognize patterns using data. These algorithms depend on statistical models and don't need to explicit step-by-step instructions. Within machine learning, supervised learning is a method where an algorithm is trained on labeled data. Once the algorithm "learns" the data, it maps inputs to outputs based on the given examples. Supervised learning comprehends of various algorithms, including regression for predicting continuous variables and classification for predicting discrete variables (Singh et al., 2016).

Ordinary Least Squares

OLS is the standard method used for linear regression models. This method works by minimizing the sum of the squares of the differences between the observed values and the values predicted by the model. Specifically, OLS finds the line that best fits the data points, that is line for which the sum of the squared errors is the minimum (Good, 2012).

Support Vector Regression

This type of algorithm is based on Support Vector Machines algorithms. The loss function in SVR allows for a margin of error, meaning that errors within a certain range are not penalized. SVR uses a kernel function to transform the input data into a higher-dimensional space where a linear regression can be estimated. The overall algorithm works by finding the hyperplane that best fits the transformed data within the specified margin. (Zhang & O'Donnell, 2020; Awad & Khanna, 2015).

Decision Tree

This method models input features and outputs in a tree-like structure. The algorithm creates a structure starting from a root node and branching out to leaf nodes, which represent output values. Decision trees resemble decision-making processes and are straightforward to understand. This helps avoid the black box issues found in artificial neural networks. (Kotsiantis, 2011; Yang, 2019).

Ensemble Learning

Ensemble learning is an approach that combines results from multiple algorithms. This method effectively reduces bias and variance, leading to more accurate models. Ensemble learning techniques include the following (Wan & Yang, 2013):

1. Bagging/Bootstrap: this technique trains multiple models using different subsets of the training dataset, typically through random sampling with replacement. The individual predictions are then aggregated through averaging or majority voting.

2. Boosting: it improves the accuracy of predictions by training models in sequence. Each model in the sequence focuses on correcting the errors made by the previous models. This process continues until a strong combined model is created, which reduces overall prediction errors.

3. Stacking: the method trains multiple models (usually of different types) in parallel and uses a meta-model to combine their predictions.

Random Forest

This is an ensemble technique that combines decision trees to reduce overfitting and improve prediction accuracy. The decision trees are trained on a random subsets of the data and features. The final prediction is made by aggregating the predictions of all the trees. (Breiman, 2001).

Gradient Boosting

Gradient Boosting is a sequential ensemble learning technique that iteratively builds models to minimize errors using the gradient of the loss function. Each new model focuses on correcting the mistakes of the previous ones, and the final model is a weighted sum of all these models, enhancing prediction accuracy. XGBoost is a version that uses parallel computing and optimization techniques to improve the algorithm performance. (Friedman, 2001; Biau et al., 2018).

Hyperparameter Optimization

Hyperparameter Optimization is the process that finds the optimal set of hyperparameters for a given model. Hyperparameters are external configurations of the model set prior to the training process, such as the learning rate in gradient descent, the number of trees in a random forest, or the kernel function in support vector regression. The most popular techniques include Grid Search, Random Search, Bayesian Optimization, and Gradient-based Optimization (Bischl et al., 2021):

1. Grid Search: evaluates all possible combinations of hyperparameters within a specified range (computationally very expensive).

2. Random Search, randomly selects hyperparameter values within specified bounds (computationally expensive).

3. Bayesian Optimization: is a method that uses probabilistic models to predict the performance of different hyperparameter sets and is computationally less expensive.

4. Gradient-based Optimization: when hyperparameters are differentiable, it uses gradient descent to find the optimal hyperparameters, providing a fast and efficient approach.

This research employs Grid Search to optimize hyperparameters for Decision Tree and Random Forest algorithms, and uses Random Search for Support Vector Regression and Gradient Boosting. Calainho et al. (2022) have shown that Random Search effectively identifies the best hyperparameter combinations for these algorithms.

Chapter 4
Results

Results

We conducted property price estimations for Atlanta, Chicago, Los Angeles, Miami, and Phoenix from 2000 to 2024. For each city, we computed 25 price estimates using algorithms such as OLS, SVR, Decision Tree, Random Forest, and Gradient Boosting. We trained models with data from each year, then predicted the log sale prices for the following year using the property features (area, property types, regions, and building area) and parameters learned from the previous year. We then calculated an index based on the estimated log price difference between periods. We analyze the Los Angeles results first due to its larger sample size. Then we conduct an inter-city comparison with the other cities to contextualize the findings. Overall, the analysis involves over 200,000 property transactions to produce more than 600 price estimates, resulting in five distinct city price indexes.

Figure 1. Los Angeles Price Index

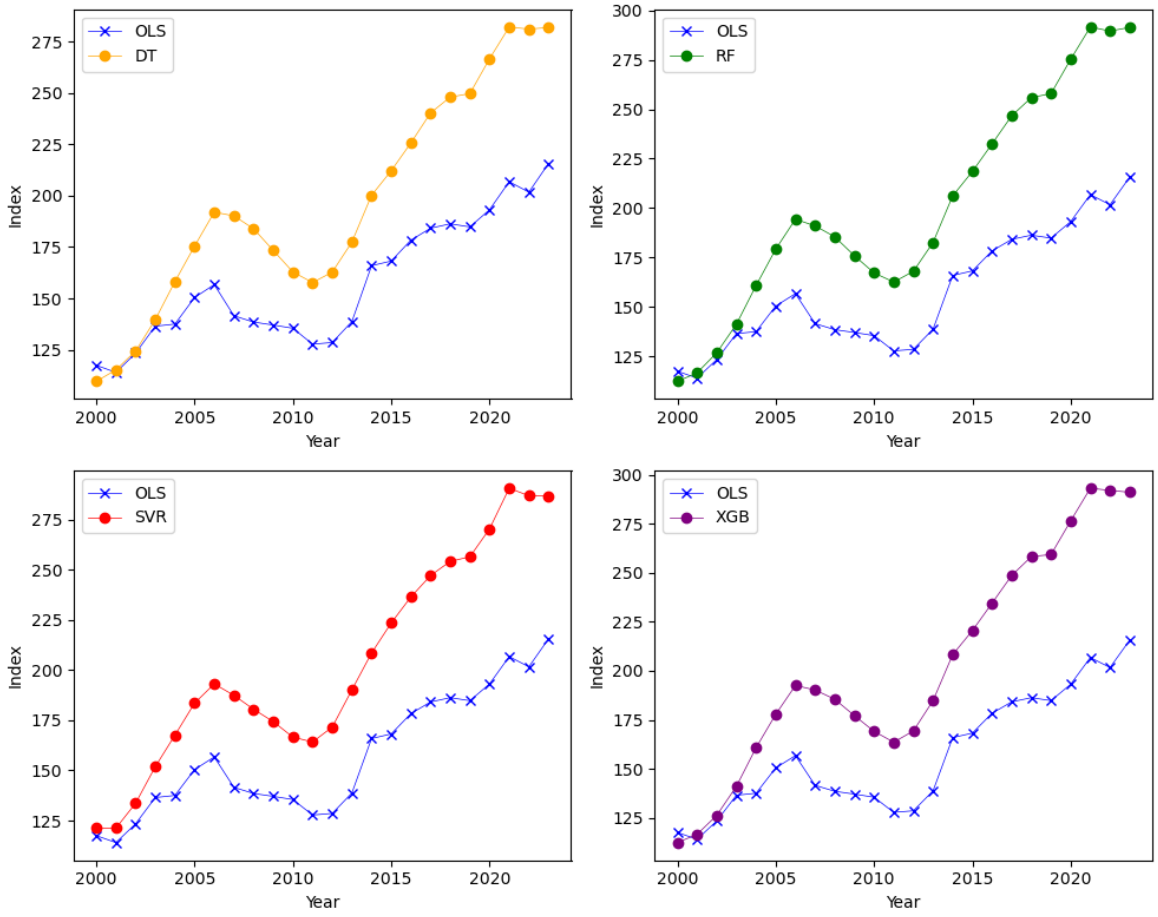


Figure 1 shows the estimated indexes from different machine learning algorithms. At first look, the index from the OLS method is more conservative, showing the lowest values among the models. On the other hand, the Random Forest model consistently records higher values, indicating it aggressively captures upward trends. Although the trends appear similar across all machine learning models at first glance, a closer look reveals noticeable differences. Specifically, the results from the Decision Tree and SVR are generally lower than those from Gradient Boosting and Random Forest. These differences might stem from how each algorithm processes data complexity and noise. Overall, these initial results suggest that our index methodology effectively provides insights into the Los Angeles property market using various algorithms.

Table 6 - Los Angeles Results Summary

Model	Returns	MSE	RMSE	Volatility	R^2
OLS	0.031217281	0.493641344	0.699411688	0.0653185	0.96163156
DT	0.043529338	0.269262305	0.518026116	0.058112549	0.857738031
RF	0.042250615	0.234284092	0.483118851	0.055192225	0.92765655
SVR	0.066795666	0.307941242	0.554140964	0.054871642	0.98278526
XGB	0.043396599	0.223051133	0.471265811	0.05511046	0.96915521

Table 6 summarizes the results from various machine learning models analyzing the Los Angeles property market. This is done by estimating the average metrics across from 2000 to 2024. The OLS model shows a return of approximately 3.12% with a relatively low volatility of 6.53% and a high R^2 value of 96%, indicating a strong fit with the observed data. The RMSE of 0.6994 suggests moderate prediction error levels. In contrast, the DT model records a higher return of 4.35% but with a significant decrease of the RMSE at 0.5180, which indicates lower prediction errors. Its R^2 value is 86%, showing a good but less accurate fit compared to OLS, and it has slightly lower volatility at 5.8%.

Looking at the models results, the Random Forest and Gradient Boosting both models yield a return of 4.22% and 6.68%, with Random Forest showing slightly better performance in terms of RMSE. Random Forest and Gradient Boosting also

demonstrate high R^2 values above 93%, indicating robust predictive accuracy. The SVR algorithm stands out with the highest R^2 value of 98% and the low RMSE values at 0.554, making it an accurate model in terms of predicting outcomes with low error and reasonable volatility at 5.5%. These models show how different machine learning approaches can variably interpret and predict market dynamics based on the same dataset.

Figure 2 shows the annual returns[7] calculated by the different models. The years 2000 and 2014 show notably high returns for OLS at 16% and 18% respectively, suggesting a strong performance in those years. SVR also peaks in 2000 with a return of 19%. The models generally show variability across the years, with DT, RF, and XGB often closely aligned in their performance metrics. The years 2007 to 2011 recorded negative returns, reflecting the global financial crisis. From 2012 onwards, there is a recovery, with all models generally showing positive returns. The SVR model frequently exhibits the highest returns, suggesting it may be more sensitive to changes in market conditions, capturing upward trends more aggressively. In contrast, OLS occasionally shows the lowest or most conservative returns, particularly in years with negative market movements, such as 2007 and 2022. This analysis highlights the diverse responses and predictive behaviors of these models over two decades in the Los Angeles property market.

These results demonstrate how different approaches predict market dynamics from the same dataset. The OLS model often provides conservative estimates, offering stability but less responsiveness to market changes. In contrast, SVR and Gradient Boosting aggressively capture market trends, beneficial in dynamic environments but also more sensitive to noise. Decision Trees and Random Forests provide a balance. This variability underscores the importance of choosing the right model.

[7] The gain or loss of the index in a particular period, represented as a percentage.

Figure 2. Los Angeles Index Returns

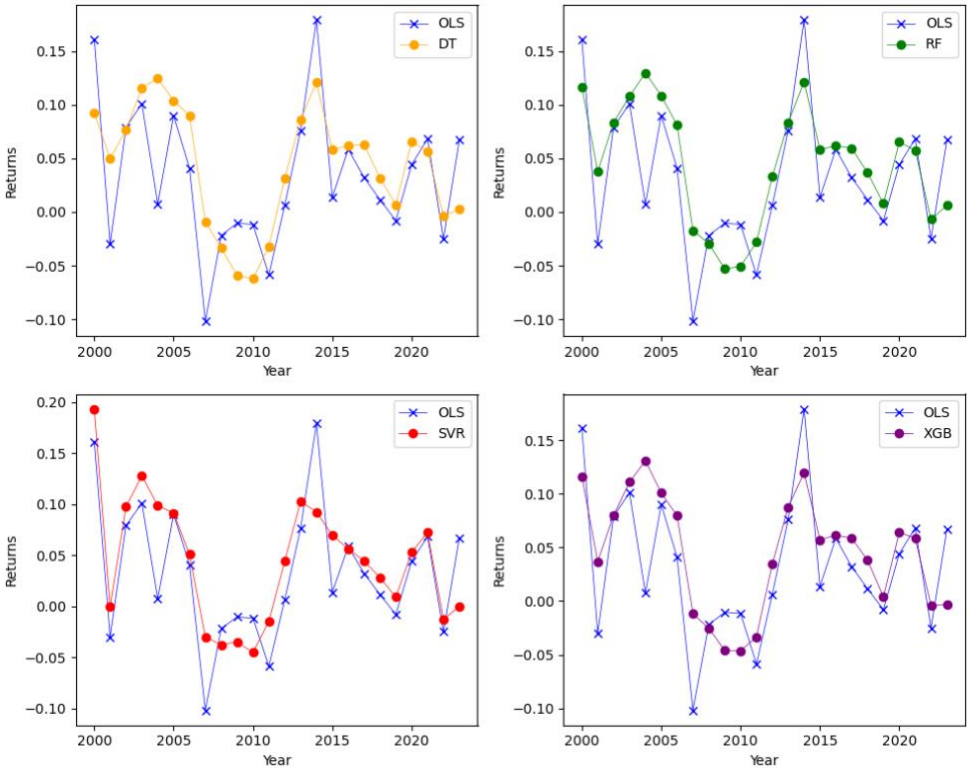


Figure 3. Los Angeles Index RMSE

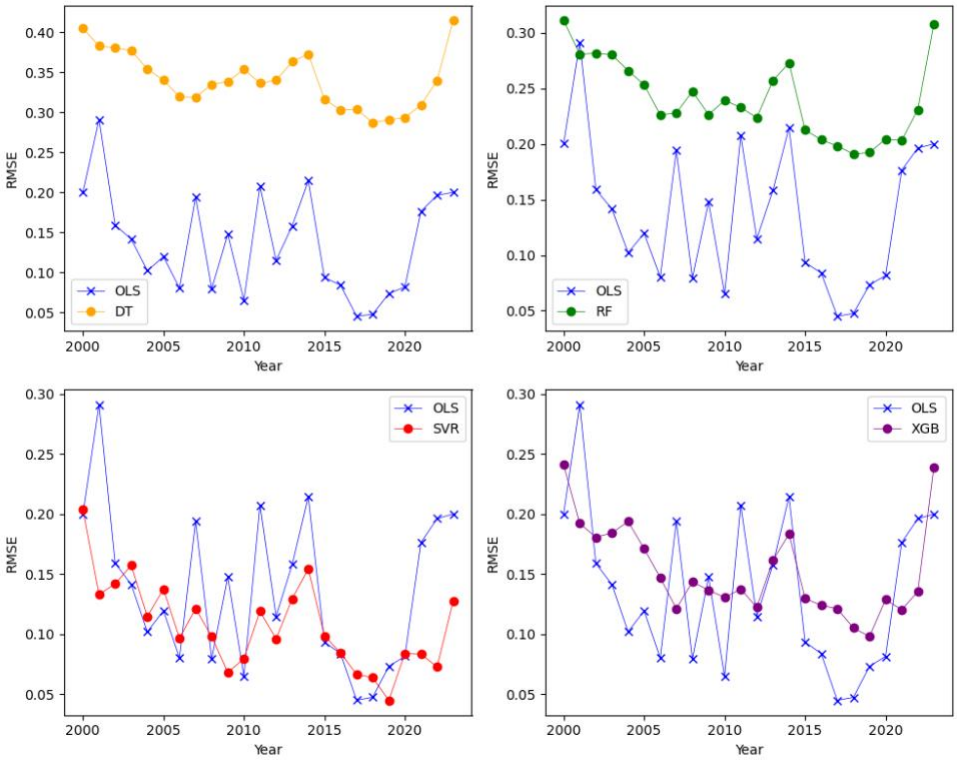


Figure 3 illustrates RMSE data to understand prediction accuracy of different models across the years. The OLS model generally shows lower RMSE values, indicating more accurate predictions, with notable exceptions in years like 2023 where it reached as high as 0.200. Decision Tree and Random Forest models typically exhibit higher RMSE values, suggesting less accuracy, with Decision Tree often having the highest.

The SVR model shows variability but generally maintains moderate RMSE levels, notably lower during the middle years such as 2019 with an RMSE of 0.045. The Gradient Boosting model displays a relatively stable RMSE, maintaining closer proximity to the SVR's performance, yet with a spike in 2023 to 0.239. Overall, these trends indicate that while some models like OLS and SVR can often provide more precise predictions, models like Decision Tree may struggle with higher errors, affecting their predictive reliability over time.

Market Comparison

In this section, we will first analyze the performance of various models across different cities, focusing on those with similar sample sizes. This approach will allow us to compare and contrast the general results and performance metrics. Following this, we will delve deeper into the accuracy of these models by comparing their performance against those from Los Angeles, which has a significantly larger sample size. This comparative analysis will help us understand how the models' accuracy and predictability vary across different market sizes and conditions.

Analyzing the Atlanta and Phoenix markets, both cities show robust model performances with some key differences. In Atlanta, the SVR is the best model with a R^2 of 96%, indicating strong predictability, and a RMSE of 66%, suggesting moderate accuracy in predictions. Similarly, Phoenix's SVR model also performs well with an adjusted R^2 of 95% and an RMSE of 11%. However, Phoenix models generally exhibit higher returns than Atlanta, with the Random Forest and Gradient Boosting models in Phoenix showing returns over 4% compared to their Atlanta counterparts, around 3%. This indicates that investments in Phoenix might yield slightly higher returns according to these models. Both cities show moderate volatility, with Phoenix slightly higher, potentially indicating more fluctuation in market values.

Comparing Miami and Chicago, two larger markets, we see distinct variations in model performance. Miami stands out with consistently higher returns across all models compared to Chicago. For instance, Miami's Decision Tree and Gradient Boosting models report returns above 5%, whereas Chicago's highest returns are around 4.3% from the Decision Tree model. Furthermore, Miami's SVR model shows a R^2 of 92.3% and a RMSE of 55.7%, suggesting accurate predictions. In contrast, Chicago's models generally show lower predictability and accuracy, with the SVR still performing best at an R^2 of 95.8% but with a higher RMSE of 71.1%. The volatility in both cities is comparable, slightly lower in Miami, indicating a more stable market. These insights suggest that Miami offers higher returns and more accurate market

predictions, making it potentially more attractive for investors seeking robust performance metrics.

Table 7 – Markets Results Summary

Atlanta

Model	Returns	MSE	RMSE	Volatility	Adj. R^2
OLS	0.02679565	0.51098779	0.7118368	0.08542131	0.94082059
DT	0.0373453	0.37227091	0.60794658	0.07769525	0.76305142
RF	0.03517515	0.31290778	0.55734956	0.07449288	0.85779383
SVR	0.03726366	0.44423174	0.66407826	0.08263405	0.95838182
XGB	0.03265631	0.2953351	0.54158684	0.07655062	0.91185959

Total = [redacted]

Phoenix

Model	Returns	MSE	RMSE	Volatility	Adj. R^2
OLS	0.02908876	0.44677925	0.66522353	0.09032266	0.94172966
DT	0.03995148	0.30689322	0.55054242	0.09656095	0.79370211
RF	0.04008096	0.25653858	0.50289835	0.09804561	0.87299059
SVR	0.04454951	0.33201681	0.57360031	0.11156964	0.949956
XGB	0.03830892	0.24092271	0.48729202	0.09671165	0.91744125

Total = [redacted]

Chicago

Model	Returns	MSE	RMSE	Volatility	Adj. R^2
OLS	0.02185162	0.64630983	0.7971649	0.09080801	0.92450745
DT	0.02269847	0.41934897	0.64455002	0.04170489	0.79067719
RF	0.02731184	0.36097169	0.59758739	0.0433456	0.87339596
SVR	0.06582769	0.51064836	0.71131671	0.08109264	0.95775276
XGB	0.02731802	0.33560303	0.57618927	0.04529161	0.93846711

Total = [redacted]

Miami

Model	Returns	MSE	RMSE	Volatility	Adj. R^2
OLS	0.04259133	0.41468054	0.63008012	0.05760519	0.94981992
DT	0.04264133	0.2521133	0.49979352	0.05710305	0.89429726
RF	0.04741716	0.26237123	0.51099309	0.06352156	0.94650006
SVR	0.06215337	0.31143192	0.55713351	0.05431678	0.92299275
XGB	0.05073711	0.28929804	0.53531747	0.07395055	0.92715705

Total = [redacted]

Figure 4. Atlanta & Phoenix RMSE

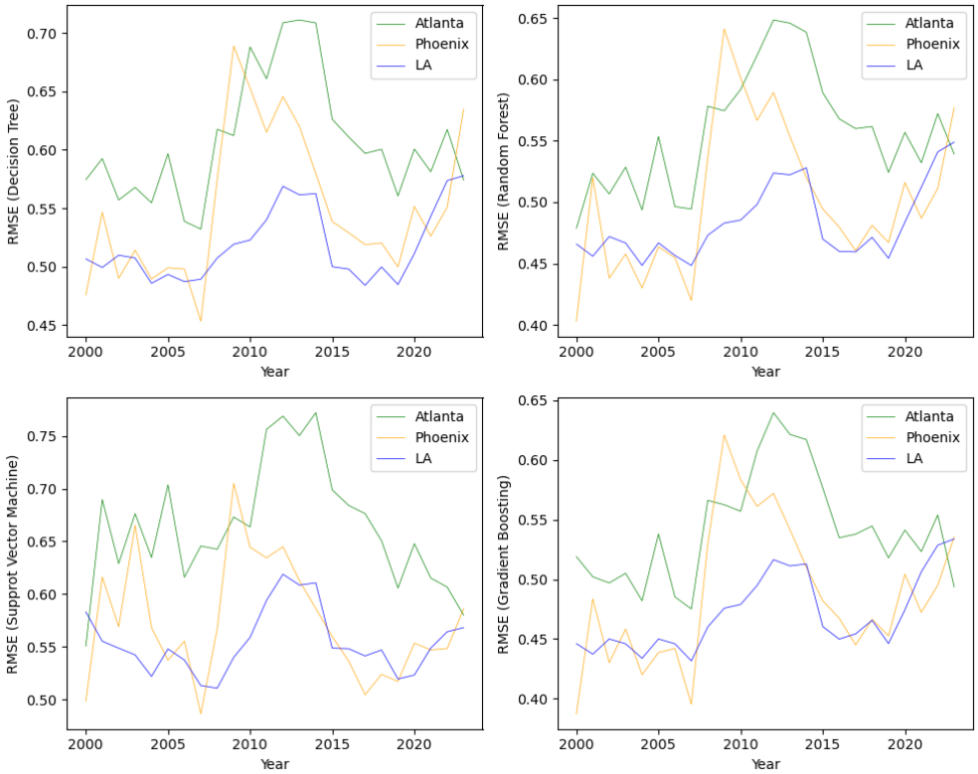
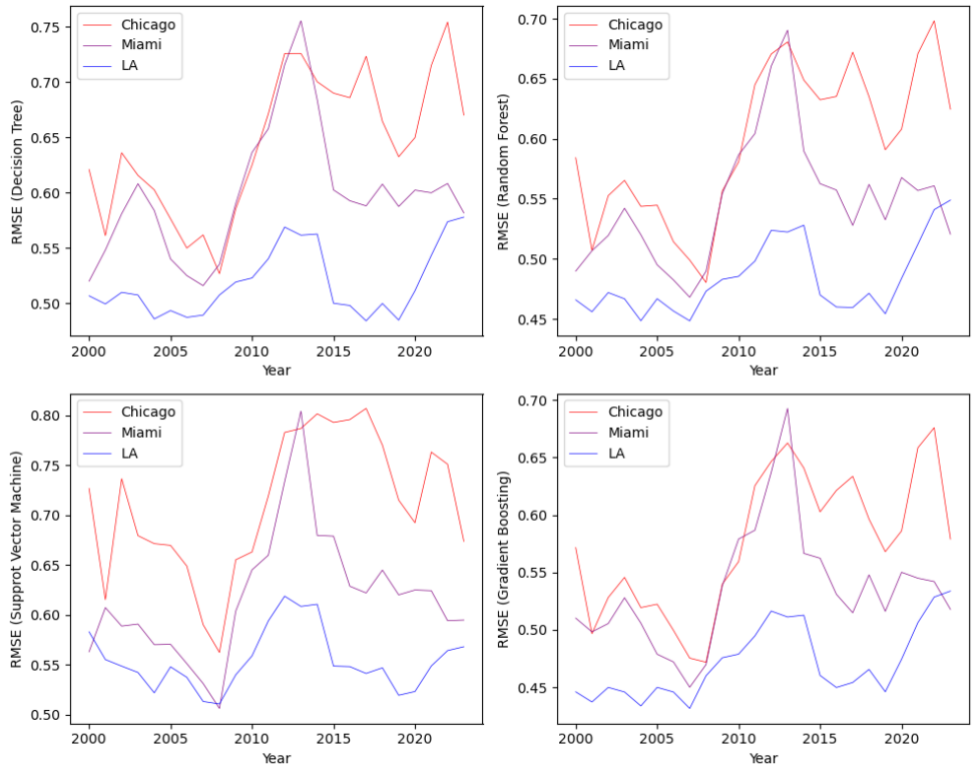


Figure 5. Chicago & Miami RMSE



Figures 4 and 5 display the RMSE results across four cities—Atlanta, Phoenix, Miami, and Chicago. It is evident that Los Angeles consistently shows lower RMSE values in most cases. This suggests that for models like Decision Tree, Random Forest, and Gradient Boosting, a larger sample size may contribute to reduced RMSE, indicating stability in the indexes However, the consistent performance of the SVR model across all cities indicates that this conclusion cannot be generalize, as SVR maintains similar accuracy regardless of sample size variations.

It's evident that there are differences in model performance that may not be solely attributed to the sample size, as some of these cities have similar sizes. Miami consistently shows higher returns and lower RMSE values, particularly in the SVR model, indicating more precise predictions and potentially higher profitability. Phoenix also exhibits robust performance with slightly higher volatility, suggesting a more dynamic market. Atlanta and Chicago, while still showing strong model performance, tend to have lower returns and higher RMSE values compared to Miami and Phoenix, indicating less accuracy in predictions.

This comparison suggests that while sample size is an important factor in model training, the intrinsic characteristics of each city's real estate market, such as economic conditions and market dynamics, also significantly influence model performance. Thus, differences in model effectiveness across these cities highlight the impact of market-specific factors over mere sample size

Table 8 – Sample Size v. RMSE Correlation

Market	Sample Size	OLS	DT	RF	SVR	XGB
Los Angeles	[redacted]	0.07983101	-0.1885481	-0.0862525	-0.2254571	0.0327374
Chicago	[redacted]	0.33761003	0.39420951	0.3432369	0.45918105	0.44550996
Miami	[redacted]	0.23238745	0.22814357	0.27220323	0.25045692	0.19437937
Phoenix	[redacted]	0.41648903	-0.0268359	0.0290601	-0.1651901	0.09102601
Atlanta	[redacted]	0.68593069	0.18364306	0.29663025	0.1879036	0.28251647

Table 5 presents the correlation between sample size and RMSE. Los Angeles, with the largest sample size of over [redacted], shows a significant negative correlation in RMSE values for some models, particularly in the Decision Tree and SVR models, indicating that increased sample size helps reduce prediction errors. However, other models, such as Random Forest, show a weaker or even positive correlation in some cases, suggesting that larger datasets may not always lead to improved performance. These correlations highlight that the impact of sample size on model accuracy varies across models, with some benefiting more from larger datasets, while others may struggle with overfitting.

In contrast, smaller markets like Chicago, Miami, Phoenix, and Atlanta display varied correlation patterns. Chicago and Miami, with slightly higher sample sizes than Phoenix and Atlanta, show a mix of negative and positive correlations, with notably lesser negative values in models compared to Los Angeles. Notably, Atlanta even shows only positive correlations, an anomaly suggesting that for certain models, increases in sample size might not lead to improved accuracy, potentially due to the underlying market dynamics or selection bias. These discrepancies underscore the complex interplay between sample size and model accuracy and highlight the importance of considering local market and data characteristics in addition to data volume.

This study, while providing valuable insights, does encounter certain limitations that suggest areas for further research. One significant oversight is the exclusion of double imputation methods that could add exhaustivity to the research. Additionally, the research does not explore the potential of using expanding and rolling window techniques, nor does it delve into further model calibration and fine-tuning. These methods are used for validating the consistency of model predictions over time, as noted by Calainho et al. (2022), who found that rolling window approaches tend to yield more accurate results.

Chapter 5
Limitations
& Conclusions

Conclusions

Limitations

This study, while providing valuable insights, does encounter certain limitations that suggest areas for further research. One significant oversight is the exclusion of double imputation methods that could add exhaustivity to the research. Additionally, the research does not explore the potential of using expanding and moving window techniques. These methods are used for validating the consistency of model predictions over time, as noted by Calainho et al. (2022), who found that rolling window approaches tend to yield more accurate results.

Another limitation is the omission of artificial neural networks from the analysis, primarily due to their high computational demands and time constraints. These models are known for their advanced capabilities, and exploration of their parameter tuning could create interesting areas of further research.

Furthermore, the research does not include stress testing, which could help to assess the robustness of models against different sample sizes. Such testing is important to ensure their reliability in real-world volatile scenarios and markets where data is not available. Addressing these gaps could significantly expand the applicability of the study's findings and provide a more comprehensive understanding of various models under different conditions.

Final Remarks

Expanding on the methodology introduced by Calainho et al. (2022), this study applies machine learning algorithms to create property price indexes across major US cities such as Atlanta, Chicago, Los Angeles, Miami, and Phoenix. The analysis, which encompasses over 200,000 property transactions, investigates the effectiveness of different machine learning models in capturing the nuances of real estate market dynamics. The comparison across these cities showcases how larger

datasets, as seen in Los Angeles, generally enhance model accuracy by providing more data, thus allowing for better generalization and fewer prediction errors.

The research findings indicate that while larger datasets typically improve the performance of machine learning models, the accuracy and reliability of these models can vary significantly between different cities. This variation is largely due to local market conditions and data characteristics, highlighting that beyond sample size, the local market characteristics are influence model performance. These insights suggest the need for tailored approaches in applying machine learning to each real estate market, considering the specific economic and market conditions of each place.

In conclusion, this research advocates for a broader integration of machine learning techniques in real estate index estimation, emphasizing that these technologies can significantly improve how property markets performance is understood. With these techniques, the industry can achieve greater precision in property valuations, enhance investment strategies, and foster a more robust understanding of market dynamics. The adoption of these technologies not only promises to elevate the operational standards of the real estate sector but also to empower stakeholders to navigate market complexities more effectively.

References

Aitkin, M., & Foxall, R. (2003). Statistical modelling of artificial neural networks using the multi-layer perceptron. *Statistics and Computing*, *13*, 227-239.

At Court. (1939). Hedonic price indexes with automotive examples.

Awad, M., Khanna, R., Awad, M., & Khanna, R. (2015). Support vector regression. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, 67-80.

Bailey, M. J., Muth, R. F., & Nourse, H. O. (1963). A regression method for real estate price index construction. Journal of the American Statistical Association, 58(304), 933-942.

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. Applied sciences, 8(11), 2321.

Biau, G., Cadre, B., & Rouvière, L. (2019). Accelerated gradient boosting. *Machine learning*, *108*, 971-992.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Lindauer, M. (2021). Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. arXiv. *arXiv preprint arXiv:2107.05847*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Calainho, F. D., van de Minne, A. M., & Francke, M. K. (2022). A machine learning approach to price indices: Applications in commercial real estate. The Journal of Real Estate Finance and Economics, 1-30.

Case, K. E., & Shiller, R. J. (1987). Prices of single-family real estate prices. New England Economic Review, Sep, 45-56.

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

Choy, L. H., & Ho, W. K. (2023). The use of machine learning in real estate research. *Land*, *12*(4), 740.

Christodoulides, C. & Christodoulides, G. (2017). The Method of Least Squares. Analysis and Presentation of Experimental Results: With Examples, Problems and Programs, 301-375.

Costa, O., & Cazassa, E. (2017). Property mix heterogeneity and market cycles: How much can we rely on median-price indices? Journal of Financial Innovation, 1(3).

DeepConverse (2024). Chatbot glossary. DeepConverse. https://blog.deepconverse.com/chatbot-glossary/

Evans, A., James, H., & Collins, A. (1992). Artificial neural networks: An application to residential valuation in the UK. University of Portsmouth, Department of Economics.

Fenwick, D. (2013). Uses of residential property price indices. In Handbook on Residential Property Price Indices, OECD Publishing, Paris.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Gatzlaff, D. H., & Haurin, D. R. (1997). Sample selection bias and repeat-sales index estimates. The Journal of Real Estate Finance and Economics, 14, 33-50.

Geltner, D., Francke, M., Shimizu, C., Fenwick, D., & Baran, D. (2017). Commercial property price indicators: sources methods and issues. Technical report, Eurostat.

Griliches, Z. (1961). Hedonic price indexes for automobiles: An econometric analysis of quality change. In Price indexes and quality change: Studies in new methods of measurement (pp. 55-87). Harvard University Press.

Gupta, D., Kose, U., Khanna, A., & Balas, V. E. (2022). Deep learning for medical applications with unique data. Academic Press.

Herath, S., & Maier, G. (2010). The hedonic price method in real estate and housing market research: A review of the literature.

Hill, R. (2011). Hedonic price indexes for housing. OECD Statistics Working Papers, No. 2011/01, OECD Publishing, Paris.

Hill, R. J., & Melser, D. (2008). Hedonic imputation and the price index problem: An application to housing. Economic Inquiry, 46(4), 593-609.

Investopedia (2023a). What an Algorithm Is and Implications for Trading. https://www.investopedia.com/terms/a/algorithm.asp

Investopedia (2023b). What Are Returns in Investing, and How Are They Measured? https://www.investopedia.com/terms/r/return.asp

Investopedia (2023c). Market Index: Definition, How Indexing Works, Types, and Examples. https://www.investopedia.com/terms/a/algorithm.asp

Investopedia (2024a). What Are Real Assets vs. Other Asset Types? https://www.investopedia.com/terms/r/realasset.asp

Investopedia (2024b). What's the difference between R-squared and adjusted R-squared? https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp

Investopedia (2024c). Real Estate: Definition, Types, How to Invest in Ithttps://www.investopedia.com/terms/r/realestate.asp

Isholafa, A. (2024). *Train, test, and dev sets.* DEV.to. https://dev.to/isholafaazele/train-test-and-dev-sets-59jh

Jurczenko, E. (2017). *Factor investing: From traditional to alternative risk premia*. Elsevier.

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.

Kolbe, J., Schulz, R., Wersing, M., & Werwatz, A. (2021). Real estate listings and their usefulness for hedonic regressions. Empirical Economics, 1-31.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*, 261-283.

Laerhoven, K., mis.tu-darmstadt.de, k., McCulloch, W., & Pitts, W. (1995). Introduction to artificial neural networks. Proceedings Electronic Technology Directions to the Year 2000, 36-62.

Liebowitz, J. (2013). Business analytics: An introduction. CRC Press.

Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR), 9(1), 381-386.

McGreal, S., Adair, A., McBurney, D., & Patterson, D. (1998). Neural networks: The prediction of residential values. Journal of Property Valuation and Investment, 16(1), 57-70.

Melser, D. (2023). Selection bias in housing price indexes: The characteristics repeat sales approach. Oxford Bulletin of Economics and Statistics, *85*(3), 623-637.

Moein, S. (2017). Definition of artificial neural network. In Artificial Intelligence: Concepts, Methodologies, Tools, and Applications (pp. 1-11). IGI Global.

Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011, June). Overview of use of decision tree algorithms in machine learning. In 2011 IEEE Control and System Graduate Research Colloquium (pp. 37-42). IEEE.

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.

Owusu-Ansah, A. (2011). A review of hedonic pricing models in housing research. Journal of International Real Estate and Construction Studies, 1(1), 19.

Persons, W. M., & Coyle, E. S. (1921). A commodity price index of business cycles. The Review of Economic Statistics, 353-369.

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. Journal of Real Estate Research, 31(2), 147–164.

Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development (INDIACom)* (pp. 1310-1315). Ieee.

Tay, D. P., & Ho, D. K. (1992). Artificial intelligence and the mass appraisal of residential apartments. Journal of Property Valuation and Investment, 10(2), 525-540.

Usher, A. P. (1931). Prices of wheat and commodity price indexes for England, 1259-1930. The Review of Economic Statistics, 103-113.

Viriato, J. C. (2019). AI and machine learning in real estate investment. *Journal of portfolio management*, *45*(7), 43-54.

Wan, S., & Yang, H. (2013, July). Comparison among methods of ensemble learning. In *2013 International Symposium on Biometrics and Security Technologies* (pp. 286-290). IEEE.

Wong, K., So, A.T., & Hung, Y. (2002). Neural network vs. hedonic price model: Appraisal of high-density condominiums. In Real estate valuation theory (pp. 181–198). Springer.

Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, *17*(1), 26-40.

Yang, F. J. (2019). An extended idea about decision trees. In *2019 international conference on computational science and computational intelligence (CSCI)* (pp. 349-354). IEEE.

Zhang, F., & O'Donnell, L. J. (2020). Support vector regression. In *Machine learning* (pp. 123-140). Academic Press.

[redacted]

[redacted]