Using Machine Learning Models to Predict Property Valuations

Aditya Sakariya (1008555225)

Abstract

This research focuses on the usage of machine learning methods to help predict property valuations. A dataset of sold, for sale and ready to build across 30,489 cities across US on 30th March 2024. This study delves into the application of machine learning (ML) algorithms to predict property valuations in the dynamic real estate market. Our analysis focused on evaluating three specific ML techniques: Random Forest, Artificail Neural Network, Extra Trees, and XGBoost. These were benchmarked against traditional hedonic pricing models to measure their accuracy in predicting property prices. The research demonstrates the potential of ML models to enhance real estate appraisals, with significant implications for the improvement of pricing strategies in the face of market changes.

1. Introduction

In the rapidly evolving landscape of real estate, technological advancements have become a cornerstone of competitive strategy among firms. The integration of digital technologies is transforming traditional practices, offering new avenues for enhancing customer experience and optimizing operational efficiency. This paper explores the pivotal role of internet-based technologies in reshaping the real estate sector, emphasizing their impact on marketing, brokerage, and property valuation services.

Historically, real estate transactions were constrained by physical boundaries and traditional marketing methods. However, the digital revolution, marked by the rise of the internet and sophisticated computational tools, has broadened the scope significantly. Today, most leading real estate firms have transitioned online, creating comprehensive platforms that not only market and sell properties but also provide enhanced decision-support tools to consumers. These platforms utilize extensive datasets, which include transaction histories, property attributes, and consumer behaviour analytics, to facilitate a more informed and efficient property valuation process.

Central to this research paper is the application of machine learning algorithms, which have shown great promise in predicting property values with high accuracy. These algorithms leverage large volumes of data to model real estate prices, taking into account various factors such as location, demographics, and market trends. This study examines the effectiveness of three specific machine learning techniques—Extra Trees, k-Nearest Neighbours, and Random Forest—in comparison to traditional hedonic pricing models, which have been the standard in property valuation.

Through quantitative analysis of transaction data and model outcomes, this research aims to provide a detailed assessment of the potential of these machine learning methods to transform real estate valuation. By doing so, it seeks to offer insights into how real estate professionals can refine their appraisal processes and enhance the precision of their pricing strategies in the face of a dynamic market landscape.

2. Literature review

Real estate plays a crucial role in the US economy by influencing various financial and economic indicators. The sector not only contributes significantly to GDP but also impacts employment and productivity. According to the National Association of Home Builders based on the quarterly seasonally adjusted (2017 dollars) analysis of Q1-2021 to Q4-2023, real estate investment (3.5%) and consumption spending on housing services (11.5%) have a combined 15% to GDP of the economy. Real estate investment is 18% of the gross private domestic investment and housing services are 16.8% of personal consumption expenditures.

Real estate markets affect consumer wealth, and their fluctuations can have profound implications on the financial health of individuals and institutions alike (Ullah & Al-turjman, 2021). Furthermore, the market's condition can influence investment decisions and economic stability. As an example, changes in real estate values can affect consumer spending through the wealth effect. Real estate investment trusts (REITs) and housing investments are directly tied to economic growth, providing clear evidence of the sector's impact on broader economic performance (Zhu & Lizieri, 2022).

Overall, real estate's integrative role encompasses not only economic aspects but also environmental and social dimensions, contributing to a comprehensive economic impact.

The real estate sector significantly influences the broader economy through various spillover effects, impacting multiple industries from construction to retail. Real estate activities drive demand in construction, significantly contributing to employment and economic activity (Gupta, Mittal, & Van Nieuwerburgh, 2022). This sector supports industries by increasing demand for raw materials like timber and cement, boosting sectors such as manufacturing and logistics. Real estate developments often lead to increased infrastructure needs, such as roads and utilities, fostering growth in public works and engineering services (Durand & Georgallis, 2018). Additionally, a thriving real estate market enhances retail sectors by increasing consumer spending through the wealth effect, where homeowners feel wealthier and spend more, which in turn stimulates the retail and services industries. The interconnections between real estate and other economic sectors demonstrate its foundational role in promoting economic growth and stability across various market segments.

Several factors influence real estate prices, including economic conditions, demographics, and government policies. Broadly, these factors are divided into physical characteristics like house features of floor area in square feet, number of bed and baths, location like the zip code or, and implicit factors like the crime rate in the area, the quality of schools in the area, availability of grocery stores in the vicinity, distance from the nearest public transport, interest rates etc.

Affordability issues arise when housing prices outpace income growth, influenced by supply constraints and fiscal measures. The predictive analysis of property valuations is crucial in assessing the risk associated with housing bubbles (Bao et al., 2022; Chou et al., 2022). In the USA, housing markets with the risk of being bubbles are closely watched due to their significant impact on the broader economy. Effective risk analysis, which includes examining historical data and market trends, helps in predicting potential market corrections and understanding the timing of housing market crashes (Ling et al., 2020; Hoesli & Malle, 2021). Policies aimed at improving housing affordability can moderate the boom-and-bust cycles characteristic of housing bubbles. Tools like machine learning models have shown promise in

providing accurate preliminary forecasts of real estate prices, which are essential for risk assessments and economic planning (Alzain et al., 2022).

Additionally, Crime rates significantly impact housing prices in an area, with higher crime often leading to lower property values. Research has shown that the perception and reality of crime can deter investment in neighborhoods, reduce desirability, and consequently depress house prices (Margaretic & Sosa, 2023). Areas with high crime rates may see a stagnation in housing price appreciation due to the increased risk perceived by potential homeowners and investors. Moreover, the type of crime—whether violent or property-related—can also influence the extent of its impact on real estate values. A detailed understanding of local crime data is crucial for buyers and sellers in the real estate market, as it affects both the immediate property values and the long-term investment potential (Akbulut-Yuksel et al., 2022). The dynamic between crime rates and housing prices underscores the importance of community safety initiatives and effective law enforcement in maintaining property values and ensuring the economic stability of neighborhoods.

The dynamics of the real estate market, including its cyclical nature and its impact on consumer wealth and spending, make understanding these factors imperative for a wide variety of stakeholders like policymakers, investors, and consumers alike. Since these dynamics are complicated and require sophisticated prediction models that minimize error. Artificial Intelligence (AI) can be deployed as a tool to develop such models using machine learning algorithms. Machine learning (ML) models provide superior performance over traditional hedonic or linear regression models in real estate valuation due to their ability to handle large, complex datasets with intricate correlations among variables. Traditional models like linear regression are limited in their capacity to capture non-linear relationships effectively, which is crucial in accurately modelling real estate prices that may be influenced by a myriad of intertwined factors (Dombrowsky, 2023; Bartlett et al., 2019).

ML models such as Random Forest and Gradient Boosting Machine (GBM) excel in their flexibility to model these non-linear and complex interactions between a property's features and its value. These models can automatically detect and model complex patterns and interactions between features without requiring explicit specification, which is a significant limitation in traditional regression models (Fan et al., 2022; Keith, 2021).

Moreover, the integration of advanced optimization algorithms like Optuna, BayesOpt, and Hyperopt for hyperparameter tuning enhances the performance of ML models. These tools efficiently fine-tune model parameters to optimize performance, which is crucial given the high dimensionality and feature interactions in real estate data (Akiba et al., 2019; Zhang et al., 2022). This capability allows ML models not only to fit the training data better but also to generalize well on unseen data, avoiding overfitting—a common problem in simpler models.

ML's application in real estate valuation extends beyond numerical data, incorporating text from listings, images from property sites, and even macroeconomic indicators, providing a holistic view of factors affecting property prices. For instance, convolutional neural networks can process and extract features from images of properties to assess aesthetic and structural elements that influence valuations (Xiao et al., 2017).

This study focuses on 4 machine learning models of Artificial Neural Networks (ANN), Extra Trees (ET), Random forests (RF), Boosting (XGBoost). These are prominent machine learning models that significantly enhance predictive analytics in various fields, including real estate

valuation. Each of these models brings unique strengths to handling complex, non-linear relationships in data, which are common in real estate markets.

Artificial Neural Networks (ANNs) are particularly advantageous due to their ability to model non-linear and complex patterns. ANNs mimic human brain operations and can process inputs through layers to capture relationships in data that linear models might miss. This capability makes them highly effective for real estate valuation, where factors such as location desirability and aesthetic qualities can be nuanced (Prieto, 2012).

The Random Forest method, combining classification and regression trees (CART) with bootstrap aggregation (also known as bagging), was initially introduced by Leo Breiman. This ensemble method enhances prediction performance by integrating several models, specifically decision trees, to form a more robust predictor. Each tree is built using a bootstrap sample of the data, which means a sample drawn with replacement from the original dataset, typically of the same size as the original. In each tree's formation, a random subset of features is selected at each node, ensuring that the trees are diverse and reducing the correlation between them (Breiman, 2001; Breiman, 1996).

This approach decreases the variance of the model without significantly increasing the bias. By averaging the results of individual trees, Random Forests mitigate the risk of overfitting associated with single decision trees, especially in cases where the dataset is noisy. The method's ability to handle large datasets with high dimensionality and its feature selection capability at each node split make it particularly effective for complex modelling tasks where interactions and non-linearities complicate the prediction landscape (Speybroeck, 2012; Berk, 2020; Choubin et al., 2019).

XGBoost, short for "eXtreme Gradient Boosting," is a highly efficient and scalable implementation of gradient boosting that has gained widespread popularity in machine learning for its performance and speed. This model is particularly effective in handling large-scale and complex data sets, making it an excellent tool for predicting real estate prices. XGBoost works by constructing a series of decision trees in a sequential manner, where each subsequent tree aims to correct the errors made by the previous ones. The method employs both L1 and L2 regularization to enhance model robustness and prevent overfitting, which is crucial in predicting real estate prices due to the noisy and diverse nature of real estate data.

Several studies have demonstrated the effectiveness of XGBoost and other machine learning models in the domain of real estate price prediction. For instance, Iwai and Hamagami (2022) developed a novel XGBoost model that incorporates boundary conditions to improve accuracy in real estate price forecasting, particularly in areas with demographic declines like Japan, addressing the inherent complexities in real estate data (Iwai & Hamagami, 2022). Similarly, a comprehensive project on Bengaluru real estate utilized XGBoost among other algorithms, finding it superior in terms of accuracy due to its robust handling of diverse data features (N., 2023).

Further, Yavuz Özalp and Akıncı (2023) compared various tree-based models, including XGBoost, for property value predictions in residential settings, underlining the high performance of XGBoost especially when robust data sets are available (Yavuz Özalp & Akıncı, 2023). The study by Nnadozie et al. (2022) implemented a multi-level stacking ensemble that included XGBoost, enhancing model accuracy significantly over traditional single-model approaches (Nnadozie, Matthias, & Bennett, 2022).

In the realm of real estate pricing, XGBoost can utilize features like location, property size, number of rooms, and additional amenities to forecast property values. This predictive power is enhanced through its ability to model non-linear relationships and interactions between features effectively. XGBoost's application in real estate pricing is supported by its ability to handle different types of data and incorporate complex business rules, which can improve investment decisions and market analysis (Yu et al., 2020; Crosby et al., 2016).

These studies collectively affirm the capability of XGBoost and related ensemble techniques to effectively predict real estate prices, leveraging their advanced computational models to accommodate the complex, multifactorial nature of real estate data.

3. Model design and methodology

This paper explores the comparative performance of Ordinary Least Squares Regressions (OLS), Extra Trees (ET), Random Forest (RF), Artificial Neural Network (ANN), and eXtreme Gradient Boosting (XGBoost) in predictive modeling tasks. These algorithms were selected for their robustness and accuracy in handling complex, high-dimensional data. First, RF, ET, and XGBoost are ensemble methods known for their capability in regression and classification tasks, with ET and RF using a multitude of decision trees to model variance effectively, while XGBoost optimizes on gradient boosting frameworks that sequentially build trees to minimize errors. ANN excels due to its deep learning framework that effectively captures nonlinear relationships in data. Second, ANN and XGBoost excel with large, complex datasets by learning intricate patterns, whereas RF and ET can handle both categorical and numerical data, making these methods adaptable to mixed data types. Third, all four algorithms are resilient against noisy data, capable of managing outliers and missing values efficiently. Fourth, their computational efficiency is notable; RF, ET, and XGBoost leverage multiple CPU cores for rapid prediction, while ANN benefits from GPU acceleration for processing large neural networks. Lastly, RF, ET, and XGBoost provide insights into feature importance, aiding in feature selection and data structure understanding, whereas ANNs offer a deep understanding of feature interactions through their hidden layers. My research will utilize RF outcomes as a benchmark to evaluate the effectiveness of ET, ANN, and XGBoost in various modeling scenarios.

Additionally, a hedonic price model is a method used in economics to estimate the market value of a product or asset based on its characteristics. It’s widely applied in real estate economics to determine the contributing value of various property attributes to the overall price. This model breaks down the item being evaluated into constituent components, each with its own assumed value.

The hedonic pricing model operates under the premise that price is determined both by internal characteristics of the goods (such as size, age, and condition of a house) and external factors (like location, neighborhood safety, and proximity to amenities). By regressing the price on these attributes, it quantifies how much each specific feature adds to or subtracts from the value. This approach is valuable for understanding how different factors influence prices in diverse markets and is also used for adjusting price indices over time, accounting for changes in product or property characteristics. This model is fundamental for researchers, policymakers, and businesses aiming to make informed decisions based on the intrinsic qualities that affect pricing.

The price (P_{it}) of a residential property (i) during time period (t) is hypothesized to be a function of a fixed number (K) of housing features, quantified as (x_{itk}). Mathematically, this relationship is described by the hedonic price model, shown in Equation (1):

$$[P_{it} = \alpha_0 + \sum_{k=1}^{K} \beta_k x_{itk} + \epsilon_{ti}]$$

Here, (α_0) represents the constant term, (β_k) denotes the coefficients associated with each housing feature, and (ϵ_{ti})is the stochastic error term. In practice, this equation is typically operationalized by regressing property prices against a range of characteristics including physical attributes, environmental factors, and accessibility features. Through this model, the implicit price of each housing attribute can be discerned, allowing researchers to accurately estimate property prices.

These methods are used to predict the house prices of sold, for sale and ready to build properties across the United States in 30,489 cities as of 30th of March 2024.

Second, the Random Forest (RF) algorithm is a supervised learning method that uses ensemble techniques for classification and regression tasks. It constructs multiple decision trees and combines them into a single predictive model, enhancing accuracy and robustness over using a single tree. The process of bagging, which involves random sampling with replacement, is employed to reduce variance while introducing a slight increase in bias. In this process, random subsets of the training set are selected repeatedly (β times), and a decision tree is fit to each subset.

Each subset corresponds to a random vector, ∅k, which uniquely influences the formation of a tree, ensuring that each tree differs slightly. The prediction for a given input X by the K-th tree is represented as $hk(X) = h(X, \emptyset k)$, where K is the total number of trees. To reduce feature correlation, each tree in the ensemble makes splits by randomly selecting features at nodes and choosing a threshold, c, that minimizes the variance in the sum of squared errors between two subsets, S1 and S2.

$$\text{SSE} = \sum_{i \in S_1} \left(v_i - \frac{1}{|S_1|} \sum_{i \in S_1} v_i \right)^2 + \sum_{i \in S_2} \left(v_i - \frac{1}{|S_2|} \sum_{i \in S_2} v_i \right)^2$$

For optimizing hyperparameters, I utilize bootstrapping and select the mean square error as my criterion. I define a range for maximum depth (1 to 20), maximum features (1 to 14), minimum samples per leaf (1 to 10), minimum samples required to split (1 to 10), and number of estimators (10 to 200). Using Optuna, I fine-tune these parameters to identify the optimal settings for my model.

Third, Extra Trees, also known as Extremely Randomized Trees, are a form of ensemble learning used for both classification and regression tasks. This method bears resemblance to Random Forest but includes notable differences in the training and integration of individual decision trees. In Extra Trees, multiple decision trees are developed using different subsets of

the training dataset, and a randomly chosen subset of features is used to decide on splits at each node of each tree. Unlike Random Forest, Extra Trees do not seek the optimal split at each node. Instead, they randomly select from potential splits that can reduce variance.

The final prediction is made by averaging the outcomes across all trees, which helps in stabilizing predictions against overfitting, as the random nature of splits ensures lower variance in individual trees. This averaging process also mitigates the impact of outliers and noise, enhancing the reliability of predictions. In contrast to Random Forest's method of generating trees from bootstrapped samples, Extra Trees use the entire training dataset for each tree and select split points randomly while sampling each feature at these points. This approach not only simplifies the splitting process but also contributes to the robustness and generalization capability of the model.

Three crucial hyperparameters are instrumental in optimizing this method. These include the number of trees in the forest (M), the number of features considered at each split (k), and the minimum number of samples required to split a node (n_min). Unlike some methods that use bootstrapping, I employ the Mean Squared Error (MSE) criterion for splitting the nodes. I establish a range for several settings: maximum tree depth (from 2 to 20), maximum features (from 2 to 14), minimum samples per leaf (from 2 to 10), and minimum samples required to split (from 2 to 10), along with the number of trees ranging from 10 to 200.

Upon applying My predictive models to the training data in Python, I obtain the optimal coefficient vector (θ) through the equation ($\theta = (X^T X)^{-1} X^T y$), which minimizes the cost for the training dataset. To evaluate the accuracy of My model on the test dataset, I input the estimated coefficients, or weights, into the model and compare predicted against actual values. I gauge the performance using mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). These performance indicators (as detailed in Equations (6) to (8)) ideally approach 0, which would indicate a perfect fit between the model's predictions and actual values.

In these equations, ($h(x^{(i)})$) denotes the predicted property value, ($y^{(i)}$) represents the true property value, and (m) is the count of observations in the test set. The equations are as follows:

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(h(x^{(i)}) - y^{(i)}\right)^2 \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2} \quad (7)$$

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|h(x^{(i)}) - y^{(i)}\right| \quad (8)$$

4. Data

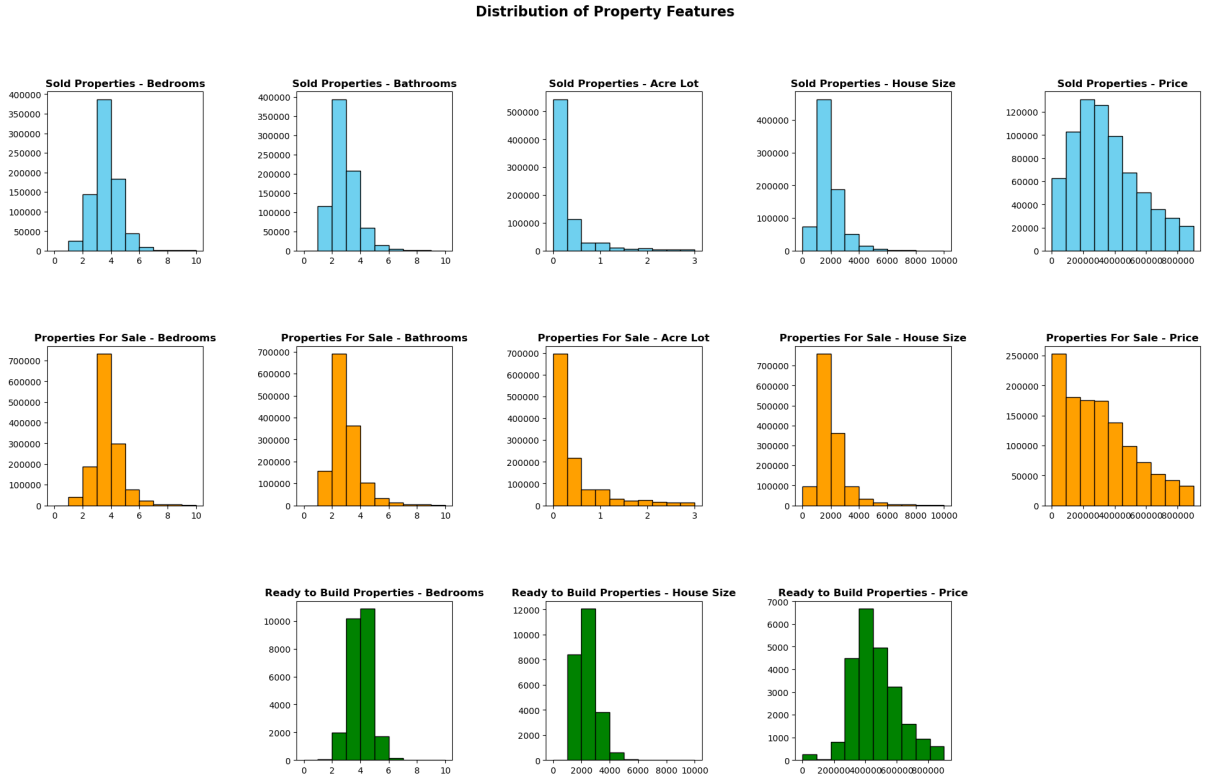
In this study, data from realtor.com was collected and then segregated based on the status of the property. This included sold, for sale and ready to build. In my research, I meticulously compiled and pre-processed three comprehensive datasets capturing various facets of the real estate market: properties for sale, ready-to-build lots, and recently sold properties. These datasets, essential to my machine learning model training, were curated from extensive realtor.com listings as of March 30, 2024, covering a broad spectrum of 30,489 cities across the United States.

The **For Sale Dataset** encompasses current market listings, providing insights into active consumer preferences and market valuations. It includes unique identifiers, detailed location information, property features such as the number of bedrooms and bathrooms, lot size, type, and age of property, alongside the current listing price. Further, it provides metadata on listing durations, pricing history, and real estate agent contacts, offering a granular view of the sales landscape.

The **Ready to Build Dataset** focuses on undeveloped land poised for construction, a segment of the market that has gained prominence with the increasing interest in custom-built homes. It details lot size, zoning information, access to utilities, and any preparatory permits or restrictions, alongside visual links to the site's current state. It's a pivotal resource for understanding the valuation of potential development projects.

The **Sold Properties Dataset** reflects completed transactions, thus serving as a historical record and a basis for predictive analytics. It extends beyond property features to include final sale prices, transaction dates, listing vs. sale price analysis, and property tax history, providing a comprehensive overview of market dynamics and property appreciation over time.
Each dataset underwent rigorous cleaning, including the normalization of features, handling of missing values, and deduplication, ensuring high-quality inputs for My analysis. The datasets were instrumental in the empirical evaluation of My machine learning models, as outlined in the results section, where the performance of OLS, RF, and XGBoost models was scrutinized across different property statuses to deduce their predictive efficacy and inform future model refinements and applications.

Figure 1:



The histograms in figure 1 offer a distribution of property features for three categories: sold properties, properties for sale, and ready to build properties. These distributions reflect the variability and central tendencies within the data for the different property features, which are key considerations for property valuation models as discussed in the research paper.

1. **Sold Properties**:
    - **Bedrooms**: The distribution is heavily skewed towards properties with fewer bedrooms, peaking at around 3-4 bedrooms. This suggests that the majority of sold properties within the dataset are likely to be family homes.
    - **Bathrooms**: Similar to bedrooms, there is a skew towards properties with fewer bathrooms, with a peak again at around 2-3 bathrooms.
    - **Acre Lot**: Most sold properties have less than 1 acre, indicating a preference or higher availability of compact land parcels in sold properties.
    - **House Size**: There is a right-skewed distribution with most houses in the smaller size range, which could reflect market availability or buyer preference for moderately sized homes.
    - **Price**: The price histogram is also right-skewed, showing that lower-priced properties are more commonly sold, which might correlate with the observed sizes and numbers of bedrooms and bathrooms.
2. **Properties for Sale**:
    - **Bedrooms**: The peak is around 3-4 bedrooms, similar to sold properties, which is typical for family homes.
    - **Bathrooms**: The trend is again similar to sold properties, with a peak at 2-3 bathrooms.
    - **Acre Lot**: There's a very steep drop-off after 1 acre, with the vast majority being under 1 acre, suggesting that larger plots are less commonly for sale or less in demand.

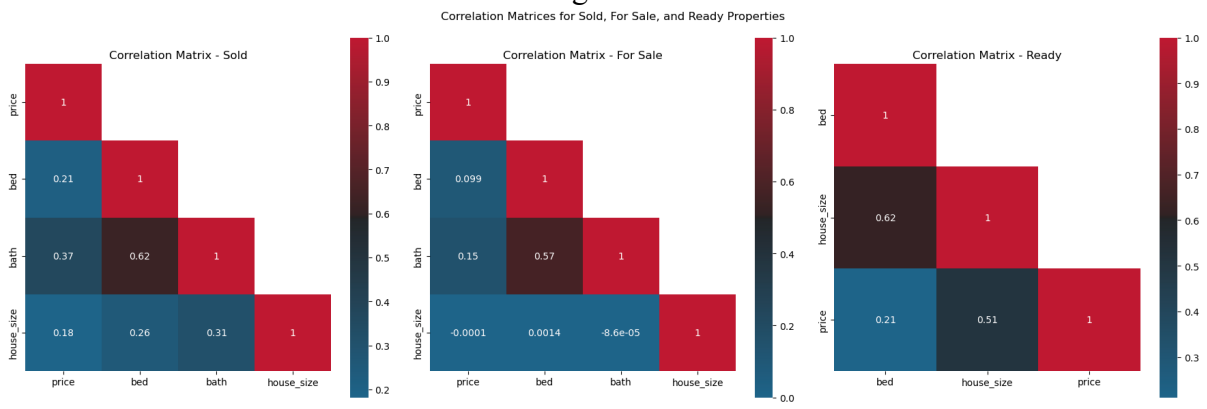
- **House Size**: The distribution is right-skewed with a large number of properties in the lower size range, which aligns with the typical house size preferences or available stock.
- **Price**: Properties for sale show a broader price distribution than sold properties, but still with a right-skew. This could indicate a wider range of property values or possibly a higher asking price compared to sold properties.

3. **Ready to Build Properties**:
   - **Bedrooms and Bathrooms**: Histograms are not provided for these categories, which is appropriate as ready to build properties might not yet have a determined number of bedrooms and bathrooms.
   - **House Size**: The distribution is more uniform with a slight right skew, indicating a more varied range of potential property sizes that are ready for construction.
   - **Price**: The distribution is again right-skewed but less steep compared to the other categories, suggesting a smaller range of prices for ready to build properties.

From the perspective of the research paper, these histograms reinforce the notion that property features such as size, bedrooms, and bathrooms are significant in determining property prices, and that there is variability within each property category that machine learning models can capture. For instance, the prevalence of certain features within the sold and for sale categories might reflect market trends that machine learning can identify and leverage for prediction.

Additionally, the right skew in price across all categories indicates that there are a few properties with very high prices compared to the majority, which might be outliers or luxury properties. Machine learning models, as discussed in the paper, are particularly adept at handling such outliers by identifying complex patterns that might not be evident from a traditional statistical approach.

Figure 2:



In the figure 2 correlation matrix visuals, there are three matrices corresponding to sold, for sale, and ready properties. Each matrix displays the correlation between different variables: price, bed, bath, and house_size.

Sold Properties Matrix: The correlation between 'bath' and 'house_size' is the strongest among the features, suggesting that the number of bathrooms is a significant indicator of the

size of the house in the sold properties dataset. The correlation between 'price' and 'bath' or 'house_size' is moderate, indicating these features have a sizable impact on the selling price of a property. This aligns with the paper's discussion that physical characteristics such as the number of beds, baths, and house size are crucial in determining property valuations.

For Sale Properties Matrix: This matrix shows a very weak correlation between the variables and the price. The 'bath' and 'house_size' again show a moderate correlation, while 'bed' seems to have almost no linear relationship with 'price'. The weak correlations may suggest that for properties that are for sale, factors not included in the matrix might play a more significant role, or the market dynamics for these properties could be different. This supports the paper's emphasis on using machine learning models that can capture complex patterns not evident in simple linear relationships.

Ready Properties Matrix: Similar to the 'Sold' category, there's a moderate to strong correlation between 'bath' and 'house_size', and these factors also have a moderate correlation with 'price'. This might indicate that for ready properties, physical attributes are good predictors of price, which is consistent with the paper's view that such features are valuable inputs for the machine learning models.
When critically analysing the results based on the paper, consider the following:

The effectiveness of ML models over traditional regression methods may be due to their ability to handle non-linear relationships and interactions between features, which could explain why 'bath' has a consistently higher correlation with 'house_size' and 'price' across all three matrices. The variance in correlation strength across different property statuses might reflect the varying market conditions and buyer preferences. Machine learning models, especially those mentioned in the paper, can adjust for such variability more dynamically.

While traditional models are limited in their predictive capacity, ML models can integrate a wide array of factors, including those not represented in the matrices, which might improve the predictive accuracy for the 'For Sale' properties where correlations are weaker.

The low R^2 values in some cases, especially for the Random Forest model in the paper's results, might indicate that many influential factors are not captured by the models or that there's a high level of inherent noise in the real estate data.

5. Results

The following section shows results obtained from RF, OLS and XGBoost Models. Models for ET and ANN were not able to be completed within the time frame of the project. They would be explored in another edition of the paper.

Figure 3:

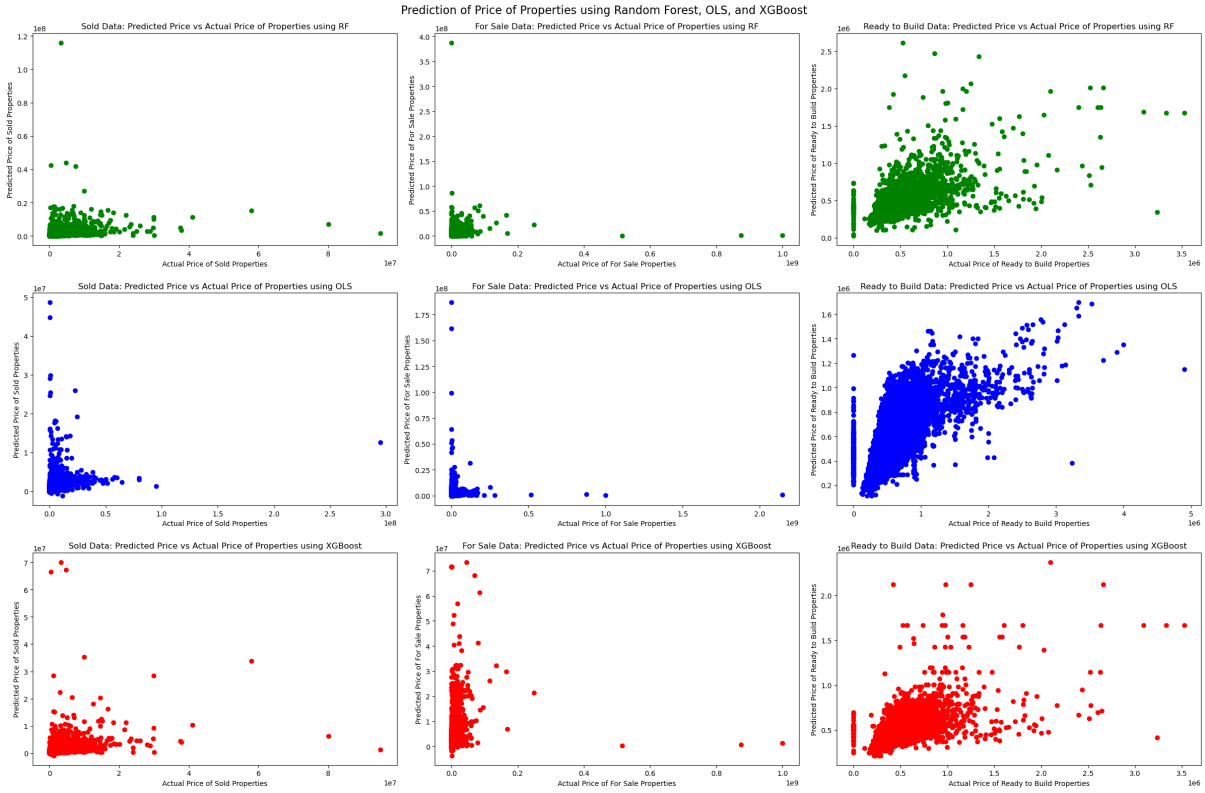


Table 1: Error Tests from OLS, RF and XGBoost models

	MAE	MSE	RMSE	R2
OLS sold properties	246402.8	4.81E+11	693750.3	0.227945
OLS for sale properties	393078.5	6.78E+12	2604698	0.032669
OLS ready to build properties	105471.6	3.13E+10	177018.3	0.515206
RF sold properties	280981.3	5.9E+11	768086	0.00139
RF for sale properties	369363.9	9.68E+12	3111568	0.001598
RF ready to build properties	115061.9	4.32E+10	207733.5	0.313872
XGBoost sold properties	262688.3	5.12E+11	716184.76	0.13178679
XGBoost for sale properties	365994.6	9.18E+12	3030560.36	0.0529064
XGBoost ready to build properties	123201.8	4.34E+10	208409.872	0.30939678

Analysing the results in table 1 and figure 3 of the models used to predict property valuations from the test date set reveals interesting insights about their performance across different

categories of real estate: sold properties, properties for sale, and ready-to-build properties. Each model — Ordinary Least Squares (OLS), Random Forest (RF), and XGBoost — presents varying degrees of accuracy and predictive power as evidenced by the MAE, MSE, RMSE, and R^2 values.

For sold properties, OLS shows an R^2 of 0.227945, which suggests that about 22.79% of the variance in the actual prices can be explained by the model. This is modest, indicating that while OLS has some predictive capabilities, a significant portion of the price variability is not captured by the model. The MAE and RMSE are also relatively high, pointing to notable average errors and volatility in predictions. RF and XGBoost models performed less effectively for sold properties, with even lower R^2 values, indicating that these ensemble methods captured less variability in the data for this category, and their predictive errors were larger.

For properties for sale, all three models showed poor R^2 scores, especially for the RF and XGBoost models, which both yielded values close to zero. This implies that almost none of the variance in sale prices was captured by these models, which suggests that there may be factors influencing the sale prices that are not included or adequately represented in the models. High MAE and RMSE values reinforce the conclusion that the predictions were not accurate for this category, which could be due to the dynamic nature of the pricing in the for-sale market or possibly due to speculative pricing strategies by sellers.

The ready-to-build properties category displayed a different trend. OLS had the highest R^2 value of 0.515206, indicating that the model was able to explain over half of the variance in the property prices. This suggests that the features used in the OLS model were relatively strong predictors for the prices of properties ready for construction. However, the RF and XGBoost models demonstrated lower R^2 values, though these were still significant compared to their performance in the for-sale category, suggesting some degree of predictive power.

The scatter plots would likely show the relationships between the predicted and actual prices for the various models. A perfect prediction model would result in a diagonal line from the bottom left to the top right of each plot. Deviations from this line would indicate prediction errors. For OLS, the predictions for sold properties and ready-to-build properties seem to have a spread that suggests a trend, whereas for sale properties the predictions are more dispersed, indicating less predictive accuracy. RF and XGBoost, as suggested by their lower R^2 values, would likely show more scatter and less trend, indicating a weaker model fit.

In evaluating the results, it's apparent that OLS was more effective in predicting prices for ready-to-build properties compared to sold and for-sale properties. This could indicate that the ready-to-build market is less volatile and more predictable based on the features used in the model, or that these properties have less variability in pricing to begin with. On the other hand, RF and XGBoost did not perform as well as expected. This could be due to overfitting, where these models are capturing noise rather than the underlying relationship, or it might be that the features and their interactions are not as influential in predicting prices as believed, necessitating further feature engineering or data collection.

While machine learning models have the potential to transform real estate valuation, their effectiveness is heavily dependent on the nature of the data and the market segment. The variability in model performance across different property categories underscores the complexity of real estate pricing and the need for models that can adapt to this complexity. Future research might explore additional features, alternative modelling techniques, or deep

learning approaches that could potentially improve predictive performance across these varied real estate categories.

6. Future implications

The future implications of this project are significant and far-reaching, considering the potential expansion of data sources and the application of diverse machine learning models. The current research provides a foundation for a more comprehensive approach to real estate valuation, and the steps outlined for future work aim to broaden and deepen this understanding.

Firstly, the proposal to scrape data for major Canadian cities like Toronto, Vancouver, and Montreal suggests an internationalization of the research. This expansion would not only provide a comparative analysis across different real estate markets but also test the robustness of the machine learning models in varied economic, cultural, and regulatory environments. Canada's real estate market, with its distinct characteristics and trends, would offer a rich dataset to further refine the models. For example, the inclusion of variables like weather patterns, which are more pronounced in Canada, or regional policies affecting housing supply, could yield new insights into the factors that drive property valuations in different geographical regions.

Furthermore, utilizing additional machine learning models such as Extra Trees (ET) and Artificial Neural Networks (ANN) can enhance the predictive accuracy of the valuation models. ET models could provide a more granular view of the impact of non-linear relationships and interactions among features due to their nature of randomizing splits and using the entire dataset. ANN models, with their deep learning capabilities, could capture complex patterns and subtle nuances in the data that simpler models might miss, such as the influence of micro-location factors or the architectural styles prevalent in a particular area.

Comparing these models' performance with the ones used in this research (RF, OLS, and XGBoost) will be instrumental in establishing a hierarchy of model effectiveness specific to real estate valuation. This could lead to the development of a hybrid approach that leverages the strengths of each model, such as using ANNs for initial feature learning followed by ET or XGBoost for prediction, thus creating a powerful ensemble model tailored for the highly variable real estate domain.

The integration of more sophisticated optimization algorithms and hyperparameter tuning methods can also be considered in future research. Advanced techniques such as genetic algorithms or neural architecture search could be employed to optimize the models further, potentially improving their generalization capabilities on unseen data.

Moreover, the expansion to different property types and market segments would allow for a more detailed segmentation analysis, which is crucial for investors, developers, and policymakers. Understanding the valuation dynamics of commercial versus residential properties, or urban versus suburban homes, can inform targeted strategies for development, investment, and regulation.

Additionally, the future work could explore the impact of emerging trends such as remote work's influence on home valuation, the significance of green building certifications, and the role of smart home technology. The ongoing evolution of the real estate market, driven by

technological innovation and changing consumer preferences, underscores the need for flexible and adaptive valuation models that can incorporate these emerging factors.

Lastly, with the advent of big data and the increasing availability of real-time data streams, future models could incorporate dynamic inputs such as current market listings, interest rates, and economic indicators, enabling more responsive and timely valuations.

This research project's future implications extend the horizon of real estate valuation significantly. By incorporating international markets, employing a wider array of machine learning models, and embracing the dynamic nature of the real estate sector, future research can provide invaluable tools for a myriad of stakeholders and contribute to a more nuanced understanding of property value determinants in a global context.

7. Conclusion

The research into the utilization of machine learning models for predicting real estate values has revealed both the complexities and the potential embedded within the real estate valuation domain. The comparative analysis of models like OLS, RF, and XGBoost has illuminated the varying degrees of effectiveness of these algorithms, with none emerging as universally superior across different property types. For sold properties, OLS showed modest explanatory power; for properties for sale, all models struggled to account for the variance in sale prices; for ready-to-build properties, OLS was notably more predictive.

These results underscore the intricacy of the real estate market and the myriad of factors influencing property values, from tangible attributes like house size to more ephemeral aspects such as consumer sentiment and market trends. While the study focused on datasets from the U.S. market, it is evident that these methodologies have global applicability and can be adapted to different market dynamics and datasets.

The research also paves the way for future studies to refine these models further. Inclusion of data from international markets like Canada, or the use of more sophisticated algorithms such as ANNs and ETs, may yield improved predictive power. There also lies the potential for real-time data integration, enhancing the models' responsiveness to market conditions.

In conclusion, this research contributes to the growing field of real estate analytics by demonstrating the practical application of ML in property valuation and providing a platform for future innovation in the sector. As we continue to refine these models and integrate more diverse data sources, we move closer to a more agile and precise valuation methodology that can keep pace with the ever-evolving real estate landscape.

References

Fahim Ullah & F. Al-turjman (2021). A conceptual framework for blockchain smart contract adoption to manage real estate deals in smart cities

Zhu, B. & Lizieri, C. (2022). Local Beta: Has Local Real Estate Market Risk Been Priced in REIT Returns?

Arpita Gupta, Vrinda Mittal, & Stijn Van Nieuwerburgh (2022). Work From Home and the Office Real Estate Apocalypse

Wensheng Bao, R. Tao, Anees Afzal, & Hazar Dördüncü (2022). Real Estate Prices, Inflation, and Health Outcomes: Evidence From Developed Economies

Jui-Sheng Chou, Dillon-Brandon Fleshman, & Dinh-Nhat Truong (2022). Comparison of machine learning models to provide preliminary forecasts of real estate prices

Elham Alzain, Ali Saleh Alshebami, Theyazn H. H. Aldhyani, & Saleh Nagi Alsubari (2022). Application of Artificial Intelligence for Predicting Real Estate Prices: The Case of Saudi Arabia

David C. Ling, Chongyu Wang, & Tingyu Zhou (2020). A First Look at the Impact of COVID-19 on Commercial Real Estate Prices: Asset-Level Evidence

Martin Hoesli & Richard Malle (2021). Commercial Real Estate Prices and Covid-19

Rodolphe Durand & Panayiotis Georgallis (2018). Differential Firm Commitment to Industries Supported by Social Movement Organizations

Mevlude Akbulut-Yuksel, N. Mocan, Semih Tumen, & Belgi Turan (2022). The Crime Effect of Refugees

Paula Margaretic & J. Sosa (2023). How Local is the Crime Effect on House Prices?

National Association of Home Builders. (n.d.). Housing's contribution to gross domestic product.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

Speybroeck, N. (2012). Classification and regression trees. International Journal of Public Health, 57(1), 243-246.

Berk, R. A. (2020). Classification and Regression Trees (CART). In Statistical Learning from a Regression Perspective.

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., & Mosavi, A. (2019). An ensemble prediction of flood susceptibility using multivariate discriminant

analysis, classification and regression trees, and support vector machines. The Science of the Total Environment, 651(Part 2), 2087-2096.

Yu, Y., Lu, J., Shen, D., & Chen, B. (2020). Research on real estate pricing methods based on data mining and machine learning. *Neural Computing and Applications, 33*, 3925-3937.

Crosby, N., Jackson, C., & Orr, A. (2016). Refining the real estate pricing model. *Journal of Property Research, 33*, 332-358.

Iwai, K., & Hamagami, T. (2022). A New XGBoost Inference with Boundary Conditions in Real Estate Price Prediction. IEEJ Transactions on Electrical and Electronic Engineering, 17.

N. (2023). Real Estate Price Prediction Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology.

Yavuz Özalp, A., & Akıncı, H. (2023). Comparison of tree-based machine learning algorithms in price prediction of residential real estate. Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi.

Nnadozie, L., Matthias, D., & Bennett, E. O. (2022). A model for Real Estate Price Prediction using Multi-Level Stacking Ensemble Technique. European Journal of Computer Science and Information Technology.