THE UNIVERSITY OF CHICAGO

Assessing the Predictive Power of Urban Green Spaces in Machine Learning Models for Chicago Housing Prices

By

Kuang Sheng

May 2025

A paper submitted in partial fulfillment of the requirements for
the Master of Arts degree in the Master of Arts in
Computational Social Science

Faculty Advisor: Crystal Bae

Preceptor: Fabricio Vasselai

Abstract

Urban Green Space (UGS) plays an important role in the urban environment. This study compares the predictive power of two types of UGS measurements — park proximity and NDVI — in the machine learning models to predict housing prices in Chicago. By incorporating real estate big data, GIS analysis, and Machine Learning techniques, the result indicates NDVI is a strong predictor for single-family residential properties on the fringe of the urban core.

Keywords: Computational Social Science; Urban Geography; GIS; Machine Learning; Real Estate

1 Introduction

Although Urban Green Space (UGS) is a fundamental part of the urban environment that generates economic benefits (Kim & Peiser, 2018), supports urban residents' physical and mental health (Russo & Cirella, 2018), and serves as a city landmark that constitutes the city image (Groos & Dages, 2008), the definition of UGS still shows significant inconsistency across different disciplines (Taylor & Hochuli, 2017). There are many great examples of centrally planned UGS implementation in global cities: Central Park in Manhattan, NY; Millennium Park in Chicago, IL; Discovery Green in Houston, TX; Ueno Park in Tokyo, JP; and Hyde Park in London, UK. At the same time, accessible UGS at a lower hierarchical level also contributes to the urban ecosystem (Gupta et al., 2016).

In academia, UGS has received significant attention from scholars in recent years due to the COVID-19 pandemic, bringing greater public awareness to the value of both physical and mental health. Researchers conducted experiments in different locations around the world to evaluate if and how interacting with UGS can alleviate stress and enhance mental wellness during the pandemic lockdown. However, most research projects solely rely on a single source of green space data: government data portals that include only government-registered park space, which are comparatively less comprehensive due to the exclusion of minor and private green spaces, like sidewalk trees, grassy medians, and private backyards, potentially leading to data deficiency and inaccuracy. Such a data collection methodology might underestimate the value of UGS in machine learning models, which have been widely used to predict housing prices in the real estate market.

At the same time, there is another type of UGS data—the Normalized Difference Vegetation Index (NDVI). NDVI is a remote sensing index that measures vegetation health and density based on remote sensing analysis of satellite imagery, providing more compre-

hensive and geographically continuous information about UGS distribution and quality. In today's real estate industry, platforms such as Zillow and Redfin have integrated machine learning-based price estimations as a key feature to enhance property valuation and user decision-making. Accurate price estimation contributes to a more transparent housing market, enhances the overall home-buying experience, and promotes social equity by reducing asymmetric information and supporting equal access to housing opportunities.

I hypothesize that compared to government data on park proximity, NDVI provides better predictive power in machine learning models. To this end, I pose the research question: *How do the two different methods of measuring UGS—distance to park space and NDVI—influence the predictive performance of machine learning models in estimating Chicago housing prices?*

For this graduate thesis, I first demonstrated the importance and significance of UGS research in social science disciplines, including geography, psychology, and urban studies, by reviewing previous studies that analyzed how UGS can shape personal experience and influence people's mental wellness based on the study of spatial cognition. Then I reviewed different methodologies that utilize the hedonic regression and machine learning models to understand the relationship between housing features and property price. Based on the theoretical framework and established real estate value estimation methodology, I explored and clarified the current knowledge gap caused by incorporating different types and levels of UGS data.

For my research, I collected Chicago listing data from Redfin (the online brokerage platform), with intrinsic housing features, and calculated environmental features using GIS software. By applying mainstream machine learning models widely used for housing price prediction and comparing the outcomes using explainable machine learning techniques, I quantified and compared the importance of intrinsic housing features and environmental features. I also evaluated the predictive power of different UGS data in different machine learning models.

2 Literature Review

Contemporary urban planning standards suggest a minimum of 9 square meters of UGS per capita, with an ideal target of 50 square meters per capita (Russo & Cirella, 2018). As an essential component of the urban environment, UGS contributes to the urban economy, provides stress alleviation, and constitutes the city's image. In this literature review, I first analyze research projects that discussed the major benefits of UGS and then approach UGS with the spatial cognition framework to understand the significance of UGS in geography, psychology, and urban studies disciplines. Finally, I conclude with a literature review of the

traditional methodology and more recent machine learning approach of UGS evaluation in real estate value prediction, which leads to the discussion of the potential data deficiency of current data collection methodology and initiates my research design of comparing the machine learning performance using different types of UGS data.

Previous researchers demonstrated the economic and social benefits of UGS. Researchers who studied Chicago's Millennium Park found that the economic benefits provided by land appreciation and business booming have greatly exceeded the investment and management of the park (Groos & Dages, 2008). Researchers also claimed a positive correlation between the interaction with the natural landscape and residents' physical and mental well-being (Bratman et al., 2012). This correlation has been recognized across different times, different geographic regions, and different academic disciplines, including public health, environmental psychology, and urban planning (Wendelboe-Nelson et al., 2019).

A study conducted in Australia found that greener spaces are contributing to both mental and physical health among adults in middle-to-older age (Astell-Burt et al., 2013). Locally, a study conducted at the University of Michigan to test and compare individual attention restoration capabilities in natural vs. urban environments demonstrates that the natural environment is more peaceful and enjoyable than any other environment. Long exposure to the natural scene or frequent interaction with the natural environment can enhance one's directed attention abilities (Berman et al., 2008). A later study investigating the association between access to urban green space and mental health found that UGS proximity and the proportion of UGS in a larger community are usually associated with decreasing anxiety (Nutsford et al., 2013). Based on this finding, more recent research studied the impact of COVID-19 on physical activity behaviors and found that outdoor physical activities can effectively alleviate anxiety (Lesser & Nienhuis, 2020). A Chicago-based research team also supported this claim by conducting comparative research in two indoor spaces: Garfield Park Conservatory (UGS) and Water Tower Place Mall (non-UGS). They found that UGS is more associated with positive feelings and creativity (Schertz et al., 2021). In conclusion, UGS is not only important but also essential in humans' daily functioning (Bertram & Rehdanz, 2015).

UGS is also significant in city image building. In many global cities, UGS provides a symbolic image as a city landmark. In different social contexts, the landmark can carry different socio-cultural meanings. Central Park has been widely accepted as a New York City landmark for decades (Rosenzweig & Blackmar, 1992). Chicago's Millennium Park was designed to be a symbolic icon that attracts tourists and boosts the local economy (Groos & Dages, 2008). In Japan's capital city, Tokyo, Hibiya Park is located right in front of the Imperial Palace, which is also close to the National Diet Building. The park's proximity to the political centers makes it an ideal location for urban rioters, who seized upon the open

spaces to rise against and overturn the top-down rule (Gordon, 1988).

All the different functionalities and benefits related to UGS can be understood in the context of spatial cognition. Geographers and psychologists have tried organizing these characteristics using Lynch's spatial cognition framework. Lynch outlined five elements of the community cognitive map in *The Image of the City*: paths, edges, nodes, districts, and landmarks. These elements constitute the methodology by which people understand the urban/community environment around them (Lynch, 1960). Based on this framework, studies have been conducted to examine the functionality and spatial cognition of UGS.

A study conducted in Beijing examined the different types of UGS in the city through cognitive mapping (Hou et al., 2021). The result showed that all five elements are presented in Beijing's UGS. What is more, different types of UGS have differences in the distribution of the respondents' cognitive maps. The differences are due to the inherent form of UGS. Landmark UGS serves as a beacon to the outside world by delivering prominent information that is easily perceived by visitors. It usually incorporates significant landscape structures and rich cultural content. What is more, a recently published paper further supported the spatial cognition of UGS by summarizing the three major components constituting one's natural experience: natural interactions, circumstances, and internal responses (Dan-Rakedzon et al., 2024). Researchers found that interacting with the outer environment can trigger internal responses. Firsthand green space experience can develop a deeper appreciation for nature.

Previous research demonstrates that UGS is significant in economic growth, mental health, and landmark construction. UGS also plays an important role in Lynch's spatial cognition framework. Although challenging to quantify directly, the value of UGS has been widely explored through housing price analyses using hedonic pricing models. In past decades, studies utilized the hedonic regression model to understand the economic effects of UGS on residential real estate. Nicholls and Crompton, who studied the housing market near greenways in Austin, TX, found a 20% value premium in properties abutting green spaces (Nicholls & Crompton, 2005). Kim and Peiser found a 25.5% value premium in properties with a view of large, passive recreational greenways in the planned community of Los Angeles, CA (Kim & Peiser, 2018). In the early 2000s, the city of Chicago launched a billion-dollar investment in implementing UGS, including Millennium Park and the Palmisano Nature Park. The successful investment made Chicago one of the world's leading cities in UGS implementation. A study of Chicago utilizing the hedonic regression model found that larger park has a positive influence on the local property price (Shaikh, 2011), while Chen et al. further extended the traditional hedonic model by incorporating the gravity model and found a positive relation between property price and the accessibility of green spaces (S. Chen et al., 2022).

In recent years, machine learning models have been widely adopted for housing value prediction and analysis in the real estate industry. Major online housing platforms have been relying on machine learning models to provide price estimation for both buyers and sellers to get a general understanding of property placement and market overview. According to Zillow's website, Zillow has strategically invested in enhancing its Zestimate, a machine-learning-powered tool that estimates home values. In the year of 2019, Zillow even launched the Zillow Prize, granting $1 million, aiming at the enhancement of the Zestimate algorithm. This initiative led to the development of the "Neural Zestimate," a deep learning model that processes data from over 100 million homes, including features like square footage, location, and images. The model achieved a national median error rate of 6.9% for off-market homes. Similarly, another major online brokerage platform player based in Seattle, WA, Redfin, also showed great interest and enthusiasm in enhancing their machine learning strategies. Redfin incorporated top tech talent, rich real estate data, and agent insights to continuously enrich the model structure and enhance the model performance. Moreover, Redfin even incorporated Multiple Listing Service (MLS) data, which is more detailed than public records, by incorporating features like whether a property has a view, is on a waterfront, or a busy street.

In addition to the practical applications of machine learning models in the real estate industry, academic interest in the field has also surged in recent years, with researchers exploring advanced algorithms to enhance predictive accuracy, pricing strategies, and market dynamics modeling. Some researchers acknowledged the shortcomings of traditional appraisal methods, which rely on hedonic pricing models. They criticized the subjectivity and potential racial bias of the methodology. A recent study addressing these concerns found that price prediction through machine learning models—particularly random forests and artificial neural networks—outperformed traditional hedonic regression in predicting housing prices in Boulder, CO, providing a more accurate and less biased alternative (Yazdani, 2021). The advantages of the Random Forest Model include: low computational cost, strong performance on small datasets, and built-in interpretability through feature importance analysis.

Meanwhile, researchers also turned to Boosting models for enhanced performance by sequentially fixing errors. For example, Gradient Boosting Model, a scalable machine learning system for end-to-end tree boosting and supports parallel computing (T. Chen & Guestrin, 2016), was used for house value appreciation prediction. Research conducted in the Greater Boston Area investigated more than 20,000 houses and found that houses with low house prices and small house areas may have a higher house appreciation potential. Their results further verified that multi-source big geo-data can be essential in machine learning frameworks to understand real estate price trends and help understand and support the

policy-making of human settlements (Kang et al., 2021). Gradient Boosting Model was selected for its high robustness and accuracy in dealing with less than clean data (Friedman, 2001). Moreover, the Categorical Boosting model (CatBoost) also shows significant advantages in house price prediction. A study conducted on housing prices in China's provincial-level regions found that CatBoost outperformed traditional algorithms with an MAPE of 12.5%, R2 of 87.81%, and EV of 90.5% (Ding et al., 2022). Besides these models, Support Vector Machine (SVM) has also been evaluated for property appraisal. A study analyzing 40,000 housing transactions over 18 years in Hong Kong found that while Random Forest and Gradient Boosting outperformed SVM in terms of MSE, RMSE, and MAPE, SVM still proved to be a valuable tool capable of producing reasonably accurate predictions under tight computational constraints (Ho et al., 2021).

To conclude, Random Forest, Gradient Boosting, and Categorical Boosting are among the most widely adopted machine learning models in real estate research, valued for their high efficiency, predictive accuracy, and ability to handle complex, high-dimensional data. These machine learning models are powerful alternative approaches to real estate evaluation, compared to traditional hedonic regression in both academic and industrial applications.

Recently, Explainable Machine Learning techniques have also received attention from the real estate and urban economics research communities due to their capabilities of facilitating researchers' interpretation of complex, nonlinear models used in property valuation. In particular, a recent study conducted in Hong Kong, China employed Explainable AI techniques such as Partial Dependence Plots (PDP), Accumulated Local Effects (ALE), and SHapley Additive exPlanations (SHAP) to dissect the influence of various housing features and neighborhood attributes on property values (Deng & Zhang, 2025). They revealed nonlinear and threshold effects for several features, such as building density, public transit accessibility, and the presence of urban amenities. As one important part of urban amenities that support urban residential well-being, the effects of UGS on housing prices using explainable machine learning techniques are worth exploring.

Despite promising results and the diversity of studies conducted on housing prices, there remains a lack of studies that compare the predictive performance of different types of UGS data. The data methodology of previous research usually only focused on municipally recorded parks and green spaces. The government's official data portals fail to include private green space and informal/small-scale green space in the city. Without considering a comprehensive analysis of UGS distribution, it would lead to an underestimated evaluation of UGS (Feltynowski et al., 2018), making UGS accessibility unequal among different social groups, and exacerbating urban segregation. In the city, private backyards, sidewalk trees, and lawns also play an important role in defining how green the neighborhood is, so in my research design and data collection, I compared green space distribution using satellite-

derived NDVI and government-reported park proximity data, applying popular machine learning models to evaluate their predictive effectiveness. Meanwhile, by leveraging housing big data, I collected the most updated housing prices and related housing characteristics as control factors.

3 Data and Methods

Drawing from the literature review, I hypothesize that relying exclusively on government-reported park proximity for UGS distribution data might be insufficient. Incorporating NDVI data collected and calculated based on remote sensing satellite imagery would provide a comprehensive assessment of UGS distribution and further enhance the predictions and performance in popular machine learning models. Consequently, my research design would involve comparing the predictive power of using governmental UGS data (park proximity) versus satellite imagery UGS data (NDVI) by testing their impact on the accuracy of multiple popular Machine Learning models for housing prices.

Firstly, I confined my data collection to the city of Chicago, IL. I picked Chicago for three major reasons:

1. Chicago is a successful city in implementing UGS. There are designed green spaces in both downtown neighborhoods (Millennium Park) and more residential neighborhoods (Palmisano Nature Park);

2. As the third most populous city in the United States, Chicago has abundant data availability in satellite imagery data and online housing big data. On December 3rd, 2023, Redfin (the online housing brokerage platform) recorded a total of 6,996 listed properties in Chicago. Although Redfin prohibits web scraping, it offers a publicly accessible download-by-batch feature that enables users to retrieve current property listing data. This open-access functionality is available to all users and reflects Redfin's commitment to data transparency. For this study, I collected the data by applying multiple price range filters to capture a comprehensive and representative dataset, while ensuring full compliance with Redfin's legal terms of service; and

3. Owing to the city's historical development, Chicago's neighborhoods exhibit pronounced diversity and segregation (Roseman et al., 1996). I anticipated that this would result in a rich and varied dataset across different neighborhoods, facilitating meaningful property comparisons based on their unique characteristics in my research.

As discussed in the literature review, the knowledge gap is associated with the granularity and accuracy of the UGS data, so the data collection process focused on two parts:

reliable sources of remote sensing data and comprehensive resources of real estate listing big data. To achieve an organized data collection, I collected the UGS data in two categories:

1. green space distribution data based on remote sensing and satellite imagery, and
2. the current boundary of municipally registered park space available on the city data portal.

For 1) green space distribution data from remote sensing and satellite imagery, I extracted the data from ChiVes: ChiVes uses harmonized, standardized environmental data at the census tract scale including the normalized difference vegetation index (NDVI). NDVI is a standardized index that calculates the greenness (biomass) based on satellite imagery. This index calculates the contrast between two bands — the chlorophyll pigment absorption in the red band and the high reflectivity of plant material in the near-infrared (NIR) band. A high NDVI value means rich vegetation in the area. A moderate NDVI value means grassland and shrubs. A very low NDVI means no vegetation, like ocean, cloud, or rocky areas. NDVI value is calculated by the following formula:

$$\text{NDVI} = \frac{\text{IR} - \text{R}}{\text{IR} + \text{R}}$$

$$\text{IR} = \text{pixel values from the infrared band}, \quad \text{R} = \text{pixel values from the red band}$$

Incorporating NDVI can avoid the data limitations of the previous research designs by expanding the definition of UGS from registered park spaces to sidewalk trees and bushes in front of the residential porch. This adjustment highlights the contrast between walking across exposed, sun-drenched environments—such as highway viaducts—and the comfort of shaded streets lined with mature trees. Such small-scale but widespread UGS can influence people's daily experience of their living environment. The enhanced data collection can fill in the gap and provide a comprehensive data source.

For 2) the current boundary of municipally registered park space in Chicago, I downloaded a ready-to-use shapefile from the Chicago Data Portal. This shapefile provides a detailed polygon representation of the officially recognized boundaries of all Chicago parks, allowing for accurate spatial analysis and mapping of municipally registered park space across Chicago. The most recent update of this layer was on October 17, 2024.

For property price and other related characteristics, I collected data from Redfin (Figure 1), a major online residential real estate brokerage platform. Redfin was founded in 2004 and recorded a 0.80% market share in the United States by the number of units sold and had about 2,000 lead agents in the year 2022. The benefit of web scraping data from Redfin is that, as a brokerage platform, Redfin pulls data and records from the Multiple Listing Service (MLS), an online service utilized by both buyers and sellers to see all homes

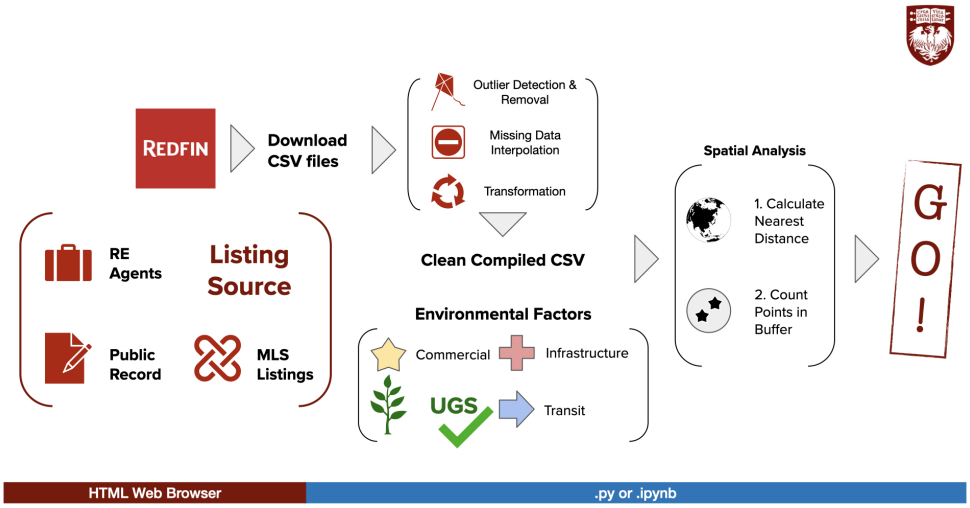

Figure 1: Redfin Data Collection Workflow

currently for sale by brokers, which ensures data accessibility and data efficiency. Redfin prohibits web scraping; however, they do allow users to download current listing data in batches for their analysis and research purposes. Redfin data provides basic and intrinsic information about the listings (like price, size, location, and number of bedrooms), which significantly enhances the efficiency and standardization of data collection. After downloading the CSV files from Redfin and proceeding with the basic data cleansing process to ensure data integrity, I enriched the dataset by georeferencing the CSV file based on latitude and longitude information and created a shapefile for geoprocessing in GIS software. I then applied vector spatial analysis (buffer) and spatial statistics to summarize the nearest distance from the listing property to the target feature and the count of feature points within the range of the buffer around the listing property. Here is a list of all the features (including geoprocessed features*) within my dataset:

- `BEDS`: Number of Bedrooms
- `BATHS`: Number of Bathrooms
- `LOT`: Lot Size
- `YR_BUILT`: Construction Year
- `DAYS ON MARKET`: Number of Days Since Listed
- `HOA/MONTH`: Monthly HOA Fees

- `LATITUDE`: Latitude of the Listing
- `LONGITUDE`: Longitude of the Listing
- `HOSPITAL_D`*: Distance to Nearest Hospital
- `RAIL_DIST`*: Distance to Nearest CTA Sta.
- `BUS_COUNTS`*: Bus Stops Count within 0.5 mile
- `FnB_COUNTS`*: Diners and Cafes within 0.5 miles (commercial ambiance)
- `PARK_DIST`*: Distance to Nearest Park
- `ndvi`*: The NDVI value of the census tract where the listing is located.
- Features with * are environmental features calculated using QGIS

For the research design, based on the locational overlay of listing data and locational features, I selected and trained machine learning models to predict listing prices with different housing-related features, wherein both the characteristics of the property itself (intrinsic) and its surrounding environment (environmental) help predict the property price. Firstly, I compared the model performance based on two different UGS datasets:

1. the Park and Green Space location on the Chicago government portal, and
2. the ChiVes NDVI dataset.

Both layers are spatial data, and I performed spatial analytics based on the location of each Redfin listing. Based on the geographic location of the listings, I quantified two types of UGS features. For governmental park data, I calculated the UGS feature based on the linear distance from the listing to the nearest park space. For the NDVI data, the UGS feature is the average NDVI of the census tract where the listing is located. These analyses were conducted in QGIS. Here is the list of operations used for spatial analysis: Buffers (vector spatial analysis) to create walkable areas around the listings, Clip (vector overlay) to extract the UGS spaces within the walkable distance, Calculate Geometry to summarize the area of UGS spaces, and spatial join to relate the listings to census tract's NDVI.

Then, I applied machine learning models to the dataset. I proposed three machine-learning models in this research design:

1. Random Forest (RF),
2. Extreme Gradient Boosting (XGBoost), and

3. Categorical Boosting (CatBoost).

According to the literature review, these models are widely utilized for their high efficiency, predictive accuracy, and ability to handle complex, high-dimensional data in real estate research. This combination of models provides a diversity of approaches and capabilities:

- Random Forest is a robust ensemble method that handles non-linear relationships and reduces overfitting through bagging;
- XGBoost is a powerful gradient-boosting framework known for its speed and accuracy, especially with structured data;
- CatBoost excels in handling categorical variables natively and requires minimal preprocessing, making it ideal for real estate datasets.

There are four model scenarios (feature configurations) in the research design to compare the predictive power of the UGS features:

1. Without any UGS data + all other housing features;
2. With UGS data from the governmental portal (Parks) + all other housing features;
3. With NDVI-enabled UGS data + all other housing features;
4. With both Park and UGS data + all other housing features.

I applied the three models discussed above to the dataset in the four model scenarios. For the model evaluation, I compiled the results into a performance comparison table, which includes Root Mean Squared Error (RMSE) and R-squared values of different models with different control groups. The performance comparison table would provide a side-by-side analysis between the models and control groups to identify whether and how different UGS data can enhance the predictive performance of mainstream machine learning models in the housing market.

Root Mean Squared Error (RMSE): RMSE is a metric used to measure the differences between predicted and actual values in a regression model. It is calculated by taking the square root of the average squared differences between the predicted and actual values. A lower RMSE means that the model's predictions are closer to the actual values, indicating optimal predictive accuracy.

R-squared ($\mathbf{R^2}$): R-squared is a statistical measure that represents the proportion of the variance in the dependent variable (housing prices) that is predictable from the independent variables (features used in the model). A larger R-squared means that a

greater portion of the variability in house prices is explained by the model, indicating a stronger fit between the model's predictions and the actual data.

I also applied explainable machine learning techniques: Partial Dependence Plots (PDP), Accumulated Local Effects (ALE), and the SHapley Additive exPlanations (SHAP) analysis to the important features from the optimal models to understand how these features relate and contribute to the house price prediction at a finer granularity.

PDP shows how a single feature affects house prices assuming feature independence. ALE provides a more nuanced view by considering the local effect of each feature while adjusting for the influence of correlated features. SHAP breaks down the impact of different features on the predicted results. I used these three methods to explain the features and understand how those features would contribute to the segregated yet diverse housing market landscape in Chicago.

In the analysis, I further applied explainable machine learning techniques to explore the effects of machine learning model prediction of UGS features (`DIST_PARK` versus `NDVI`) based on different property types of Chicago. I explored how residents with different preferences of property type evaluate UGS.

This research design addresses a critical gap in the existing literature—the overreliance on government-reported park data as a proxy for UGS—by incorporating the NDVI value derived from satellite imagery. NDVI offers a continuous measure of vegetation coverage, capturing finer-scale UGS elements such as tree-lined sidewalks and residential greenery that are often overlooked in park-based metrics.

Meanwhile, the housing big data was sourced from real Redfin listings, which ensured data abundance and accuracy. Furthermore, Redfin listing data was complemented by georeferencing and spatial analysis in GIS software, ensuring the dataset is both accurate and rich in contextual detail.

To ensure methodological rigor, the study applied three of the most widely chosen machine learning models in housing price prediction: Random Forest, XGBoost, and CatBoost. These models were selected due to their high capability of handling complex, nonlinear relationships and avoiding overfitting. To further validate the model performance, the study also employed two standard evaluation metrics—RMSE and R^2—which collectively measure prediction accuracy and the explanatory power of each model.

Besides predictive performance, the study further enhanced interpretability using explainable machine learning tools, including PDP, ALE, and SHAP analysis. These tools uncover the contribution of each variable to predicted outcomes, offering insight into the mechanisms behind housing prices and further examining and comparing the effects of the two types of UGS data.

This study presents a data-driven approach to evaluating how different types of UGS

data influence housing price predictions in machine learning tasks. Focusing on Chicago, the methodology combined real estate big data, spatial analytics, and machine learning techniques to compare the model's predictive performance and feature importance. The results aimed to examine if incorporating more granular and vegetation-based UGS data would offer measurable improvements to model accuracy and interpretability, which contributes to a better understanding of both real estate analytics and urban planning research.

4 Results

4.1 Model Performance Comparison

In this study, I utilized Python and advanced machine learning libraries—including Scikit-learn, CatBoost, and XGBoost—to implement and evaluate three high-performing models: CatBoost, Random Forest, and XGBoost, applied to a geospatially enriched dataset of residential property listings in Chicago. The dataset was analyzed in its entirety, as well as across different subsets of property types (Condo/Co-op or Single Family Residential), with a particular focus on controlling UGS features (No UGS, `DIST_PARK`, `NDVI`, or Both). Model performance was assessed using RMSE and R-squared metrics to evaluate the accuracy and explanatory power of each model in predicting housing prices.

Here are the Results (Figure 2):

4.1.1 Result comparison between models

Overall, CatBoost shows the most stability and robustness in predicting housing prices among the three tested models. When applied to the entire dataset, in the case of No UGS, CatBoost shows the lowest RMSE (103593) and the highest R-squared (0.844), significantly outperforming the Random Forest (RMSE = 109,712, R^2 = 0.825) and XGBoost (RMSE = 106,039, R^2 = 0.836) models. Moreover, CatBoost delivers more optimal outcomes than the other two models in almost all subsets of property types, all configurations of UGS feature inclusions, and both RMSE and R-squared measurements, which suggests the stability, adaptability, and robustness of the CatBoost model to varying data category, feature selection, and evaluation metric. More specifically, in the single-family residential subtype, CatBoost shows significant advantages over the other models, especially when incorporating the `NDVI` feature (RMSE = 108,975, R^2 = 0.831). Results indicate that the CatBoost model, among the three tested models, can best capture the non-linear relationship in the housing value prediction and deliver the most stable prediction outcomes in the real estate contexts.

Property Types	Models					
All	**CatBoost**		**Random Forest**		**XGBoost**	
	RMSE	R2	RMSE	R2	RMSE	R2
No UGS	103593	0.84395	109712	0.82497	106039	0.83649
DIST_PARK	106973	0.8336	110321	0.82302	108614	0.82845
NDVI	104382	0.84156	110056	0.82387	107241	0.83276
Both	109886	0.82441	110175	0.82349	111441	0.81941
Condo/Co-op	**CatBoost**		**Random Forest**		**XGBoost**	
	RMSE	R2	RMSE	R2	RMSE	R2
No UGS	113590	0.82869	113712	0.82832	114936	0.8246
DIST_PARK	111721	0.83428	113876	0.82782	113916	0.8277
NDVI	111872	0.83383	114387	0.82628	113005	0.83045
Both	113640	0.82854	114126	0.82707	112939	0.83065
Single Family	**CatBoost**		**Random Forest**		**XGBoost**	
	RMSE	R2	RMSE	R2	RMSE	R2
No UGS	113020	0.81861	113337	0.81759	119275	0.79798
DIST_PARK	111476	0.82353	113212	0.818	121491	0.7904
NDVI	108975	0.83136	112650	0.8198	116642	0.8068
Both	109414	0.83	112750	0.81948	120858	0.79258

Figure 2: Model Performance Comparison

4.1.2 Results controlling different UGS features

For the entire dataset (all property types), all three models deliver the most accurate predictions when none of the UGS features are included, which indicates the potential overfitting problem of UGS features in the machine learning tasks. However, `NDVI` does show slightly better outcomes than `DIST_PARK` (CatBoost: RMSE-2,591, R^2+0.008), indicating low enhancement. Including both `NDVI` and `DIST_PARK` features delivers the least optimal outcomes (CatBoost RMSE = 109,886, R^2 = 0.824), which suggests that information redundancy might increase feature complexity and cause overfitting when analyzing all the property types in the Chicago dataset.

For the Condo/Co-op subtype, the change between different UGS feature configurations is minimal. Moreover, the relative importance of `NDVI` and `DIST_PARK` features varies across the three models, making it difficult to conclude or summarize which features constantly contribute to the model prediction among the condo property types. Such patterns illustrate the complexity of features influencing the condo value. Other intrinsic housing features (like floors and size) or environmental features (like commercial ambiance and transportation connectivity) might show higher feature importance in the model. Neither type of UGS shows significant predicting power in the machine-learning tasks of condo properties.

For the Single Family Residential subtype, the `NDVI` feature significantly enhances the

results of all three models. Especially in CatBoost and XGBoost models, NDVI enhances the model performance by R^2 +0.00783 / RMSE -2,501 (CatBoost: Moderate Improvement) and R^2 +0.01640/RMSE -4,849 (XGBoost: Moderate to Strong Improvement) compared to the DIST_PARK feature. In addition, compared to no UGS feature, NDVI improves performance by R^2 +0.01275 / RMSE -4,045, indicating a moderate enhancement. To conclude, NDVI consistently yields better predictive performance than no UGS feature or DIST_PARK feature. The enhancement ranges from minor (Random Forest) to moderate/strong (CatBoost and XGBoost), suggesting that in the context of single-family residential property type, NDVI, as a comprehensive UGS indicator that measures both quality and quantity of the green space, can better capture relevant variation in housing values than DIST_PARK, a simple linear proximity measure indicating the straight-line distance from the listings to UGS.

4.1.3 Results between different property types

The entire dataset contains the largest training dataset and tends to show the best predictive performance in almost all models. The best performance is All - No UGS in CatBoost, which delivers an RMSE of 103,593 and an R^2 of 0.844.

In the single-family residential subset, the NDVI feature shows different levels of performance enhancement in all three models, with the most significant enhancement in CatBoost. However, UGS-related features show no improvement in the model performance when analyzing the Condo/Co-op property type, which might be explained by the complexity of the condominium's locations and the associated pricing strategies.

To conclude, across all model configurations and data subsets, CatBoost consistently outperformed Random Forest and XGBoost in terms of both RMSE and R^2. Among UGS features, NDVI showed the most stable contribution to predictive accuracy, particularly for single-family homes, whereas the combined use of NDVI and DIST_PARK did not yield further improvement and, in some cases, decreased model performance. Notably, the sensitivity to UGS variables varied by property type, with single-family homes being most responsive, and Condo/Co-op units showing marginal effects.

4.2 Explainable Machine Learning

Based on the results of the model performance comparison, I continued to apply explainable machine learning techniques to explain the influence of both UGS features (DIST_PARK and NDVI) and other significantly important features based on the CatBoost Model (model with the most optimal performance) and two property type groups (All listings and Single-Family Residential) that constantly show optimal outcomes.

4.2.1 Permutation Feature Importance

	feature	cb_perm_mean	cb_perm_std
14	FnB_COUNTS	[illegible]40.7843	6040.6961
3	BATHS	[illegible]373.9562	6652.1246
8	LATITUDE	[illegible]4510.1807	6208.5564
5	YR_BUILT	27375.9498	3069.4172
2	BEDS	26896.0552	5552.0365
7	HOA/MONTH	11853.4044	1818.6185
9	LONGITUDE	10191.3911	776.9479
0	index	8946.2578	1507.9061
4	LOT	8840.9706	3021.2758
12	RAIL_DIST	8743.6729	1780.0241
1	ZIP	7274.2703	377.4013
15	ndvi	5382.3018	1265.5874
10	PARK_DIST	3426.5349	1110.6242
13	BUS_COUNTS	2849.0128	882.4602
6	DAYS ON MA	2578.1625	836.2341
11	HOSPITAL_D	854.3829	1641.4332

	feature	cb_perm_mean	cb_perm_std
3	BATHS	[illegible]99.5433	6931.6652
8	LATITUDE	[illegible]479.9027	5764.9478
14	FnB_COUNTS	41797.9273	6396.4352
5	YR_BUILT	20548.7763	5692.9507
11	HOSPITAL_D	7415.0196	2974.2423
12	RAIL_DIST	6014.4241	3273.5933
6	DAYS ON MA	5387.8213	872.6564
15	ndvi	5157.7313	1504.4986
4	LOT	4727.5802	2828.7261
9	LONGITUDE	4465.7765	2471.4496
2	BEDS	3946.6161	1632.9438
13	BUS_COUNTS	3654.7916	1027.1732
0	index	3142.2635	1171.9300
10	PARK_DIST	3067.0646	1077.9583
1	ZIP	180.3727	1673.8727
7	HOA/MONTH	0.0000	0.0000

	feature	cb_perm_mean	cb_perm_std
3	BATHS	[illegible]425.1007	6735.4680
5	YR_BUILT	[illegible]081.3415	6113.1247
14	FnB_COUNTS	37081.3411	4446.5427
2	BEDS	14988.4946	2599.8008
8	LATITUDE	11932.0003	5631.6102
0	index	11482.5600	1967.1001
1	ZIP	6748.1354	2294.3537
7	HOA/MONTH	4744.1234	3303.5789
9	LONGITUDE	3969.8619	902.7591
6	DAYS ON MA	2481.6105	1445.9822
13	BUS_COUNTS	1967.8604	1980.9507
11	HOSPITAL_D	1509.4921	1804.2138
15	ndvi	1096.6852	1552.4728
10	PARK_DIST	1046.8998	1797.7278
12	RAIL_DIST	894.2394	841.4780
4	LOT	0.0000	0.0000

Figure 3: Permutation Feature Importance (Left: All, Middle: Single-Family, Right: Condo/Co-op)

Figure 3 shows the permutation feature importance of the CatBoost Model (labeled as cb in the list). The results are grouped by property type: the left column shows outcomes for all property types, the middle column focuses on single-family residential properties, and the right column presents results for condo/co-op properties. Both intrinsic housing features (`BATHS`, `BEDS`, `YR_BUILT`) and environmental features (`FnB_COUNTS`, `LATITUDE`) are among the top of the lists.

Features that show great importance in all property type groups are: `FnB_COUNTS`, `BATHS`, and `LATITUDE`. `FnB_COUNTS` suggests that commercial ambiance around the housing significantly affects housing desirability, reflecting a preference for vibrant, amenity-rich community areas. Among intrinsic features, `BATHS` stands out as a strong predictor across all property types. The number of bathrooms often reflects the overall size and amenity level of a property—larger homes with more bathrooms typically indicate a higher standard of living and are priced accordingly. `LATITUDE` also plays a key role due to Chicago's well-defined north-south axis and significant north-south segregation. Property values tend to shift notably across latitude lines. For instance, northern neighborhoods such as Lincoln Park often command higher prices compared to areas further south, due to differences in perceived safety, infrastructure, and investment.

Focusing on UGS-related features, the model results highlight a moderate yet consistent influence. `NDVI`, which measures the surrounding greenery, was ranked 12th in importance when considering all property types, and rose to 8th when limited to single-family residential properties. This upward shift indicates that greenness and vegetative cover hold greater relevance for single-family homes. In contrast, distance to the nearest park (`DIST_PARK`) ranked 13th overall and 14th for single-family homes. However, in the condo/co-op group,

both NDVI and DIST_PARK ranked lower than in the single-family category, indicating that urban greenness and park proximity may play a less significant role in condo pricing. While both UGS features function as support variables in price prediction, NDVI remains the stronger and more consistent predictor, particularly for single-family homes.

4.2.2 Partial Dependence Plot (PDP)

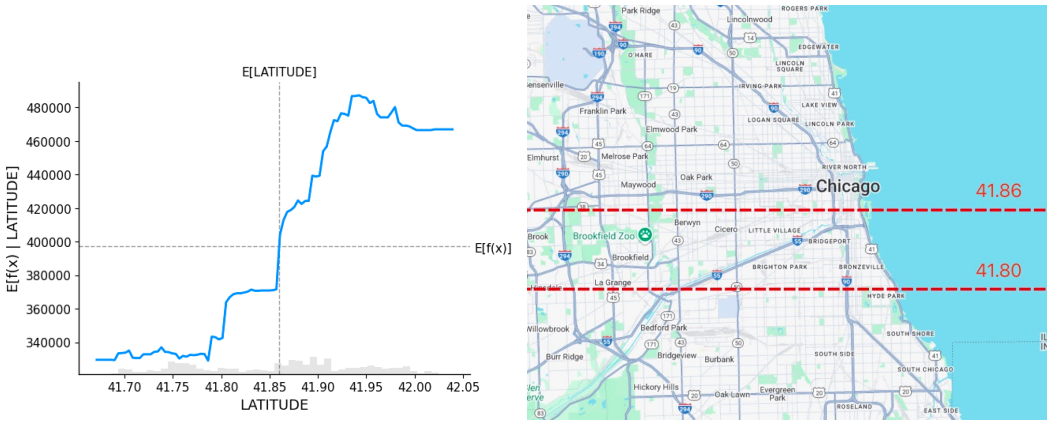

Figure 4: Left: PDP of LATITUDE; Right: Map of Chicago Neighborhoods and Key Latitude

Figure 4 shows a significant jump at Latitude 41.80, indicating a key shift in housing values. What is more, the CatBoost model further highlights another sharp increase at Latitude 41.86. After comparing the latitudes on a Chicago map, those locations correspond to the community areas of Hyde Park and Downtown Chicago (South Loop). Moving toward the northern neighborhoods, as the latitude continues to increase, the predicted value also goes up steadily, reflecting a shift into the more expensive community areas of Near North Side and Lincoln Park. This pattern reinforces a pronounced north-south price divide in Chicago's housing park and highlights the significance of Hyde Park in Chicago's south.

Figure 5 shows the PDP of UGS features:

i) `NDVI` – **All Property Types:** The plot shows a relatively stable trend between 0.1 and 0.2, typically corresponding to downtown areas with limited vegetation. Starting from `NDVI = 0.2`, property values begin to decline sharply, reaching their lowest point around `NDVI = 0.4`. Beyond 0.4, as properties move away from the downtown core, values begin to recover, stabilizing around $402,500.

ii) `NDVI` – **Single-Family Residential:** This curve starts higher overall, indicating that green surroundings are more valued in this segment. There is a sharp drop between

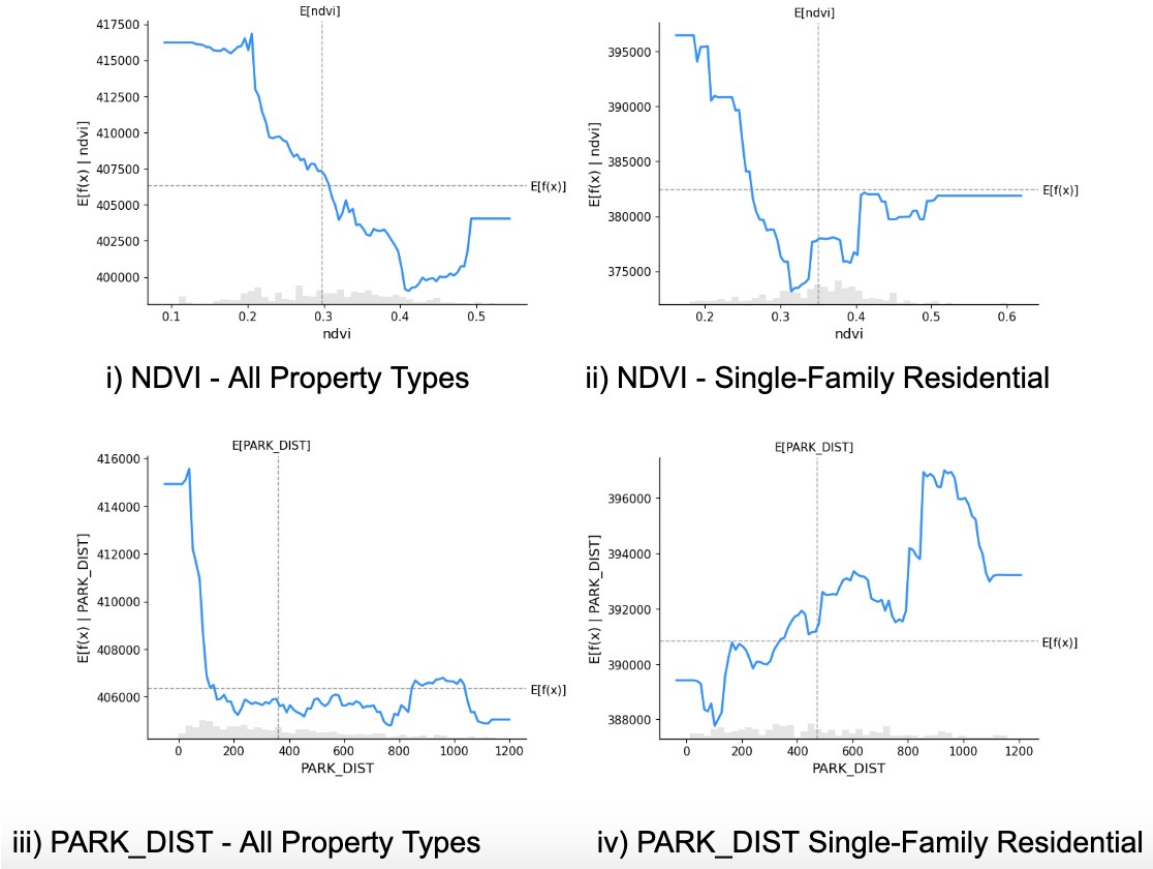

Figure 5: PDP of UGS Features

0.2 and 0.3, followed by a gradual upward trend from 0.3 to 0.6, eventually stabilizing—suggesting that suburban greenness positively influences single-family home prices.

iii) `PARK_DIST` – **All Property Types:** The highest predicted price occurs at zero distance to parks, but drops sharply within the first 200 meters. Beyond that, from 200 to 1200 meters, the curve levels off, indicating that proximity beyond a certain threshold doesn't significantly affect the price.

iv) `PARK_DIST` – **Single-Family Residential:** This plot shows a general upward trend, with prices increasing from around $388,000 to $396,000 as the distance from parks increases. While the line has small fluctuations, the overall trend suggests that being slightly farther from parks may be more desirable. This observation appears counterintuitive when compared to existing literature, which often highlights proximity to parks as a positive amenity. A possible explanation may involve concerns related to privacy, foot traffic, or noise levels near park areas. These alternative interpretations

will be explored further in the discussion section.

4.2.3 2D Accumulated Local Effects (ALE)

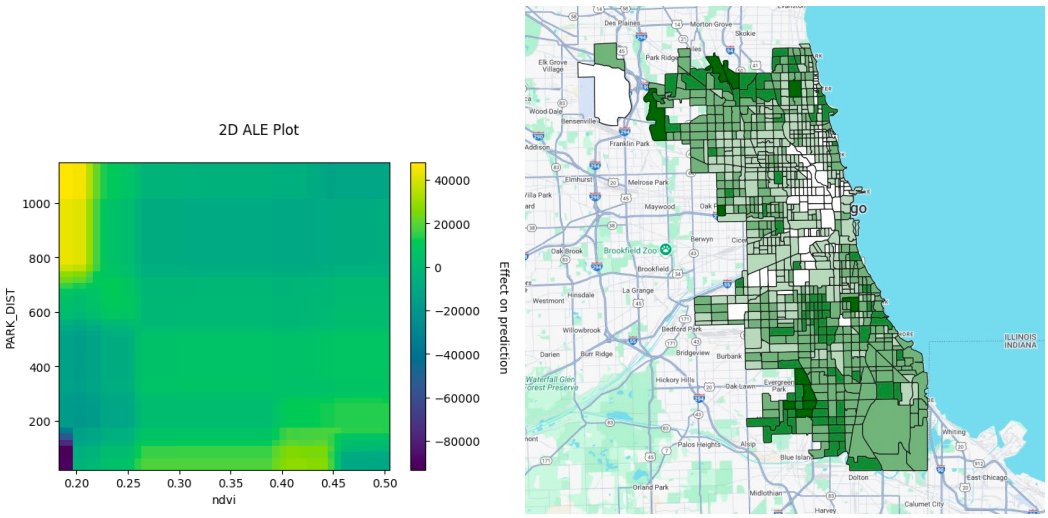

Figure 6: Left: 2D ALE of `NDVI` vs. `PARK_DIST` for Single-Family Residential; Right: Map of `NDVI` Value (5 Categories)

Figure 6 illustrates the interaction between `NDVI` and `PARK_DIST` for single-family homes in the dataset. In the lower-left corner (low `NDVI` neighborhood, close to the park), which likely corresponds to dense central urban areas, there is a strong negative effect (around −$80,000) on predicted property prices. This suggests that, in the city center, being close to a park doesn't necessarily add value—possibly due to insecurity of public space or lack of privacy. Conversely, in the upper-left quadrant (high `NDVI` neighborhood, far from the park), and the lower-right quadrant (high `NDVI` neighborhood, close to the park), there are positive value spikes (+$40,000). Those locations include central but quieter locations that are away from public park traffic and suburban or fringe locations with better vegetation and easy park access—conditions often preferred by single-family homebuyers. The rest of the surface is relatively flat (ALE near 0), indicating neutral interaction effects across typical areas.

4.2.4 SHAP Analysis

Figure 7 presents SHAP values for single-family residential, condo/co-op, and all property types, respectively, offering a visual interpretation of the relative importance and effect of all other key features in the housing price prediction models. Among all the property type combinations, `NDVI` consistently ranks higher than `DIST_PARK` in SHAP value importance,

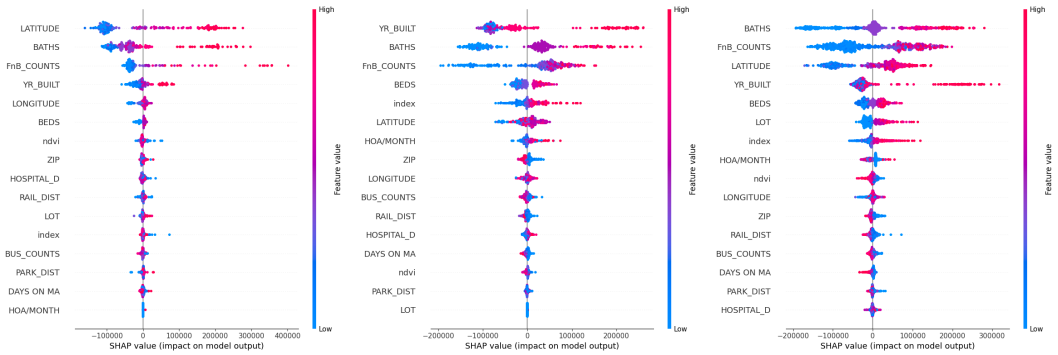

Figure 7: SHAP Value Plots (Left: Single-Family Residential, Middle: Condo/Co-op, Right: All Property Types)

indicating that `NDVI` derived from satellite imagery exerts a stronger influence on model predictions than proximity to government-designated parks. However, the rank difference between `NDVI` and `DIST_PARK` is more pronounced in the single-family residential group, where `NDVI` climbs significantly in importance while `DIST_PARK` drops lower in the feature hierarchy. In contrast, the condo/co-op group shows a much narrower gap between the two, with both features occupying relatively lower positions overall. This suggests that urban greenness plays a more prominent role in valuing single-family homes, possibly due to greater private land exposure and outdoor space utilization, while its influence is less emphasized in the more compact, amenity-driven condo market.

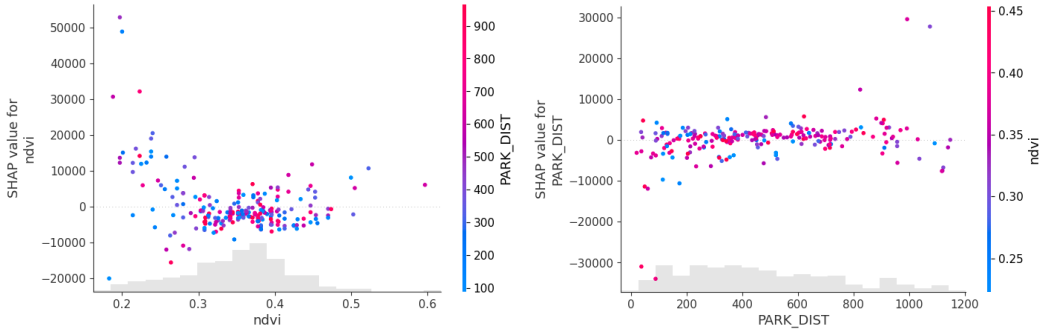

Figure 8: Left: SHAP Scatter Plot for `NDVI`; Right: SHAP Scatter Plot for `PARK_DIST` (Single-Family Residential)

To further investigate the marginal effects of UGS features in the single-family residential case where the rank difference is more pronounced, SHAP scatter plots were generated for `NDVI` and `PARK_DIST`. In the plot on the left (Figure 8), where SHAP values of `NDVI` are colored by `PARK_DIST`, a clear nonlinear trend emerges: lower `NDVI` values are associated with decreasing SHAP contributions between 0.2 and 0.3, which corresponds with the city

center area (Figure 6), while SHAP values gradually rise and stabilize as NDVI exceeds approximately 0.3. Notably, many of the higher NDVI values are colored red, indicating they correspond to greater distances from parks. This pattern suggests that private or non-park-related vegetation (captured by NDVI) exerts a positive influence on housing prices, even in areas farther from public parks.

In contrast, the plot on the right (Figure 8), which depicts SHAP values of PARK_DIST colored by NDVI, shows little consistent trend, with most SHAP values clustering around zero. The color distribution appears mixed across the PARK_DIST spectrum, implying that proximity to parks does not have a stable or significant effect on price predictions. Together, these visualizations reinforce that NDVI—an indicator of surrounding greenness—plays a stronger and more reliable role in shaping housing values than the simple distance to the nearest park.

In summary, the permutation feature importance results demonstrate the importance of both intrinsic housing features (such as BATHS) and environmental characteristics (particularly FnB_COUNTS and LATITUDE) in predicting property values. FnB_COUNTS emerged as a top environmental factor, emphasizing the value of vibrant, amenity-rich neighborhoods. When focusing on UGS features, NDVI consistently demonstrated greater predictive strength than PARK_DIST, especially for single-family residential properties. This suggests that the overall quality and quantity of UGS around a home may be more influential in shaping property values than simple proximity to parks. In SHAP value importance, NDVI consistently outperforms PARK_DIST, highlighting its stronger and more consistent influence on housing prices across property types.

5 Discussion

These results illustrate the predictive contribution of UGS indicators—specifically NDVI and distance to the nearest park (DIST_PARK)—to housing value models across multiple property types in Chicago. The results offer several insights into the nuanced role of UGS in housing markets and machine learning-based valuation tasks in highly urbanized areas.

Surprisingly, across all tested machine learning models, the inclusion of UGS features did not universally enhance predictive accuracy. For the entire dataset, models generally performed best in the absence of UGS features, suggesting a potential overfitting effect or noise introduced by green space variables. This reveals the complexity of the urban environment and highlights the importance and significance of commercial ambiance and intrinsic housing features in highly urbanized areas.

However, among the tested UGS indicators, NDVI consistently outperformed DIST_PARK in terms of improving model performance—particularly in the case of single-family res-

idential properties. For instance, CatBoost and XGBoost models using `NDVI` exhibited moderate to strong improvements in RMSE and R^2 compared to those using `DIST_PARK`, indicating that vegetation density and quality may offer a more meaningful representation of perceived environmental value than simple linear proximity to green space. Conversely, in Condo/Co-op properties, no consistent predictive advantage was observed for either UGS variable. The varying influence of UGS across property types is among the most salient findings of this study. Single-family residential properties exhibited the greatest sensitivity to UGS indicators, especially to `NDVI`. This pattern aligns with urban planning literature suggesting that suburban and low-density residential buyers tend to value natural aesthetics, environmental quality, and private outdoor space more heavily in their location preferences. In contrast, Condo/Co-op units showed marginal sensitivity to UGS variables, suggesting that other attributes—such as proximity to public transportation, commercial centers, or accumulation of workspace—likely play a larger role in shaping price variation in dense urban contexts. Observing the distribution of Condo/Co-op vs. Single-Family Residential (Figure 9) also reveals that Condo unit listings are mostly accumulated around the center of the city, while Single-Family Residential listings show more scattered locational patterns, which echoes the point that downtown condo units are more likely to be influenced by amenity-rich urban features.

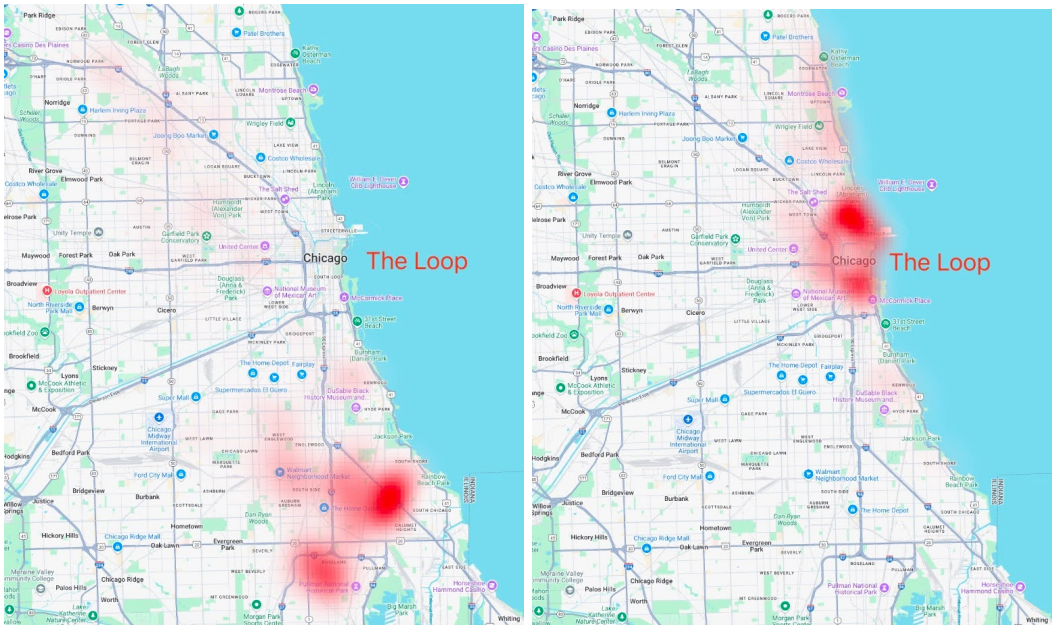

Figure 9: Heat Map: Property Type Distribution (Left: Single-Family Residential, Right: Condo/Co-op)

From a modeling perspective, among the different tested models, CatBoost shows the

most stable and robust performance, outperforming Random Forest and XGBoost across almost all data subsets and feature configurations. Notably, CatBoost retained predictive advantage even when controlling for additional UGS features, affirming its suitability for capturing complicated and nonlinear relationships in urban environments.

Permutation feature importance further confirmed the moderate but consistent predictive role of `NDVI`, particularly in single-family models where it ranked 8th in overall importance. `DIST_PARK`, by contrast, ranked lower in all scenarios, suggesting that park proximity alone may be less objective and influential in UGS's prediction on housing values.

Based on the PDP results, the influence of UGS features on housing prices varies notably by property type and indicator. For `NDVI`, a nonlinear relationship is observed across all properties: prices decline with moderate greenness (`NDVI` $\approx$ 0.2–0.4), then recover as greenness increases beyond 0.4, possibly reflecting a transition from dense urban areas to greener suburbs. This effect is more pronounced for single-family homes, where the initial price level is higher and the positive impact of higher `NDVI` is stronger and more sustained, supporting the idea that suburban greenness is more valued.

Interestingly, for single-family homes, predicted prices tend to increase with greater distance from parks, suggesting that close proximity to parks may not always be perceived as advantageous. A closer examination of Chicago's park data reveals that many parks in South Chicago are equipped with active recreational facilities such as basketball courts and baseball fields. In contrast, many passive green spaces, golf courses, and cemeteries commonly found in the northern areas are not included in the park dataset used for analysis. This discrepancy may help explain the counterintuitive trend and points to a potential preference among buyers for quieter, more private environments. These findings suggest that `NDVI` and `PARK_DIST` capture distinct dimensions of UGS value, with `NDVI` demonstrating greater consistency and contextual relevance, especially in the single-family housing market.

The 2D ALE plot further explored the relation between `NDVI` and park proximity and revealed their complex spatial interactions. In dense urban cores (low `NDVI` neighborhood, close to park), predicted housing values declined significantly, indicating that parks in such contexts may not serve as amenities but potentially as liabilities—possibly due to crowding, noise, or safety concerns. Conversely, locations with either high `NDVI` or moderate park distance showed neutral to positive valuation effects, reinforcing the conclusions in the reviewed literature that passive green space, instead of municipally recorded parks, is more desirable for residents.

SHAP value analysis results further confirm `NDVI` as a stronger predictor of housing prices than `DIST_PARK`, especially for single-family homes where high `NDVI` aligns with higher property values. Even in models covering all property types, `NDVI` remains more influential despite spatial dilution.

However, there are several limitations in this research. Firstly, the granularity of `NDVI` data is limited by the data source. `NDVI` is collected and calculated based on census tracts of Chicago. Compared to building-level calculation, this method is more efficient, but less accurate considering that different listings might have different green space interactions even within the same census tract. Moreover, we must realize the limitations of NDVI in urban settings. Although NDVI provides continuous information about UGS based on satellite imagery, NDVI often works better with a larger geographic region where vegetation coverage is extensive. The presence of buildings and shadows in urban environments may reduce the reliability of NDVI, potentially underestimating the availability and quality of UGS. Secondly, the research was conducted in a highly urbanized environment—Chicago, which may limit the generalizability of the findings. The complexity of multiple urban features, high urban density, and vegetation distribution patterns differ significantly between cities, suburban, and rural areas. Thirdly, the data were collected at a specific time point, which introduces a limitation related to temporal variability. Real estate markets demonstrate significant temporal features. A temporally continuous data collection methodology can enhance the robustness of the research design.

6 Conclusion

This study examines and highlights the sophisticated yet significant role that UGS indicators, specifically NDVI and distance to parks, play in housing price machine learning models in Chicago. While the inclusion of UGS features might not consistently enhance model performance across all property types, NDVI emerged as a particularly strong predictor for single-family residential properties, which are scattered around in more residential neighborhoods on the fringe of the urban core, underscoring the quality and quantity of UGS in determining perceived environmental standard. In contrast, the influence of UGS on Condo/Co-op properties in the city center was less pronounced, indicating that other characteristics and urban amenities related to urban convenience and connectivity—such as proximity to public transportation and commercial facilities—may be more influential and outplay UGS's effects in those high-density urban neighborhoods, especially in the city center. The findings also suggest that machine learning models, particularly CatBoost, offer effective methods to capture complex and non-linear relationships between housing prices and UGS.

This study offers insights for researchers in both the real estate industry and the public policy sector. It provides a foundation for enhancing house price prediction models based on the availability of features, supporting more informed investment strategies, resolving asymmetric market information, and offering guidance for policy-making aimed at improv-

ing social equity through real estate interventions and public infrastructure planning.

Based on the findings, urban planners and developers should tailor green infrastructure strategies based on property type and location—prioritizing high-quality, accessible UGS in suburban areas where single-family homes prevail, while focusing on incorporating UGS with transit-oriented and highly commercialized development in dense urban cores where UGS has less impact. Real estate professionals can use NDVI metrics as an additional source to guide site selection and marketing, especially for environmentally sensitive buyers. Meanwhile, data scientists should develop specific machine learning models for the type of property and explore interactions between UGS and urban amenities to better capture localized effects on housing prices.

Data and Code Availability Statement

All code used for data processing, analysis, and model development in this study is available on Github. The Redfin housing data used in this project was publicly downloaded following Redfin's terms of service and is not redistributed in the repository. Environmental data derived from publicly available sources, including NDVI data from ChiVes and park shapefiles from the City of Chicago Data Portal, are referenced in the repository with links for access. Instructions for data acquisition and replication of results are included in the README.

References

Astell-Burt, T., Feng, X., & Kolt, G. S. (2013). Mental health benefits of neighbourhood green space are stronger among physically active adults in middle-to-older age: Evidence from 260,061 Australians. *Preventive Medicine*, *57*(5), 601–606. https://doi.org/10.1016/j.ypmed.2013.08.017

Berman, M. G., Jonides, J., & Kaplan, S. (2008). The Cognitive Benefits of Interacting With Nature. *Psychological Science*, *19*(12), 1207–1212. https://doi.org/10.1111/j.1467-9280.2008.02225.x

Bertram, C., & Rehdanz, K. (2015). The role of urban green space for human well-being. *Ecological Economics*, *120*, 139–152. https://doi.org/10.1016/j.ecolecon.2015.10.013

Bratman, G. N., Hamilton, J. P., & Daily, G. C. (2012). The impacts of nature experience on human cognitive function and mental health. *Annals of the New York Academy of Sciences*, *1249*(1), 118–136. https://doi.org/10.1111/j.1749-6632.2011.06400.x

Chen, S., Zhang, L., Huang, Y., Wilson, B., Mosey, G., & Deal, B. (2022). Spatial impacts of multimodal accessibility to green spaces on housing price in Cook County, Illinois. *Urban Forestry & Urban Greening*, *67*, 127370. https://doi.org/10.1016/j.ufug.2021.127370

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Dan-Rakedzon, N., Fleming, W., Lissovsky, N., Clayton, S., & Shwartz, A. (2024). A framework for understanding the human experience of nature through cognitive mapping. *Conservation Biology*, e14283. https://doi.org/10.1111/cobi.14283

Deng, L., & Zhang, X. (2025). Boosting the accuracy of property valuation with ensemble learning and explainable artificial intelligence: The case of Hong Kong. *The Annals of Regional Science*, *74*(1), 32. https://doi.org/10.1007/s00168-025-01365-7

Ding, X., Wang, W., Zhang, Y., & Zhong, X. (2022). Machine Learning-based Models for House Price Prediction in Provincial Administrative Regions of China: https://doi.org/10.2991/aebmr.k.220307.035

Feltynowski, M., Kronenberg, J., Bergier, T., Kabisch, N., Łaszkiewicz, E., & Strohbach, M. W. (2018). Challenges of urban green space management in the face of using inadequate data. *Urban Forestry & Urban Greening*, *31*, 56–66. https://doi.org/10.1016/j.ufug.2017.12.003

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Gordon, A. (1988). The Crowd and Politics in Imperial Japan: Tokyo 1905-1918. *Past & Present 121*.

Groos, N., & Dages, M. (2008). Millennium Park: A Model for Successful Urban Green Space Redevelopment.

Gupta, K., Roy, A., Luthra, K., Maithani, S., & Mahavir. (2016). GIS based analysis for assessing the accessibility at hierarchical levels of urban green spaces. *Urban Forestry & Urban Greening*, *18*, 198–211. https://doi.org/10.1016/j.ufug.2016.06.005

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

Hou, Y., Qu, Y., Zhao, Z., Shen, J., & Wen, Y. (2021). Residents' Spatial Image Perception of Urban Green Space through Cognitive Mapping: The Case of Beijing, China. *Forests*, *12*(12), 1614. https://doi.org/10.3390/f12121614

Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, *111*, 104919. https://doi.org/10.1016/j.landusepol.2020.104919

Kim, S. K., & Peiser, R. B. (2018). THE ECONOMIC EFFECTS OF GREEN SPACES IN PLANNED AND UNPLANNED COMMUNITIES. *Journal of Architectural and Planning Research*.

Lesser, I. A., & Nienhuis, C. P. (2020). The Impact of COVID-19 on Physical Activity Behavior and Well-Being of Canadians. *International Journal of Environmental Research and Public Health*, *17*(11), 3899. https://doi.org/10.3390/ijerph17113899

Lynch, K. (1960). *The Image of the City*. Cambridge: MIT Press.

Nicholls, S., & Crompton, J. L. (2005). The Impact of Greenways on Property Values: Evidence from Austin, Texas. *Journal of Leisure Research*, *37*(3), 321–341. https://doi.org/10.1080/00222216.2005.11950056

Nutsford, D., Pearson, A., & Kingham, S. (2013). An ecological study investigating the association between access to urban green space and mental health. *Public Health*, *127*(11), 1005–1011. https://doi.org/10.1016/j.puhe.2013.08.016

Rosenzweig, R., & Blackmar, E. (1992). *The park and the people: A history of Central Park*. Cornell University Press.

Russo, A., & Cirella, G. (2018). Modern Compact Cities: How Much Greenery Do We Need? *International Journal of Environmental Research and Public Health*, *15*(10), 2180. https://doi.org/10.3390/ijerph15102180

Schertz, K. E., Saxon, J., Cardenas-Iniguez, C., Bettencourt, L. M. A., Ding, Y., Hoffmann, H., & Berman, M. G. (2021). Neighborhood street activity and greenspace usage

uniquely contribute to predicting crime. *npj Urban Sustainability*, *1*(1), 19. https://doi.org/10.1038/s42949-020-00005-7

Shaikh, S. L. (2011). The economic impact of urban green space investments: A case study for Chicago. *Program on Global Environment and Public Policy Studies, University of Chicago*.

Taylor, L., & Hochuli, D. F. (2017). Defining greenspace: Multiple uses across multiple disciplines. *Landscape and Urban Planning*, *158*, 25–38. https://doi.org/10.1016/j.landurbplan.2016.09.024

Wendelboe-Nelson, C., Kelly, S., Kennedy, M., & Cherrie, J. (2019). A Scoping Review Mapping Research on Green Space and Associated Mental Health Benefits. *International Journal of Environmental Research and Public Health*, *16*(12), 2081. https://doi.org/10.3390/ijerph16122081

Yazdani, M. (2021, October). Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction [arXiv:2110.07151 [econ]]. https://doi.org/10.48550/arXiv.2110.07151