Information Systems Student Research Journal

Research Paper DOI: ISSN:

Machine Learning for Real Estate Price Prediction

Nektarios Hajiyianni, Vasso Stylianou
Department of Computer Science, School of Sciences and Engineering
University of Nicosia, Cyprus
hajiyianni.n@live.unic.ac.cy

Abstract:

The real estate sector faces persistent challenges in achieving accurate and timely property valuations due to complex market dynamics, regional variability, and fluctuating economic conditions. Traditional valuation methods often fall short in capturing these factors, leading to pricing inefficiencies and increased investment risk. This research explores the transformative potential of machine learning in the real estate sector, focusing specifically on property price prediction. By leveraging advanced modeling techniques and integrating both internal and external factors, the research demonstrates the potential for more accurate property valuations. The study compares different predictive models, ranging from basic prediction models to more advanced approaches like CatBoost. The aim is to enhance price prediction accuracy by leveraging Gradient Boosting models and time-series forecasting techniques to capture the nuances of real estate pricing. The findings have significant practical applications for real estate investors, developers, and property managers, from optimizing portfolio decisions to assessing financial viability and navigating market downturns. The models presented equip stakeholders with robust tools for informed, data-driven decision-making, highlighting the value of machine learning in fostering a more resilient and adaptive property market.

Keywords: Machine Learning, Real Estate, CatBoost

This manuscript underwent peer review. Sune Dueholm Müller served as the managing editor.

1 Introduction

The real estate sector is being transformed by technological advancements, particularly in artificial intelligence and data science. This study explores the intersection of machine learning (ML) and real estate, highlighting how these technologies improve decision-making. As a subset of artificial intelligence, ML involves algorithms and statistical models that enable computers to perform tasks without explicit instructions, learning from data and improving their performance over time. Data mining, as defined by Han et al. (2012), involves uncovering patterns, correlations, and trends in large datasets using statistical methods and algorithms.

Real estate professionals often struggle to interpret large, fragmented datasets, making it difficult to set accurate prices, identify trends, or respond quickly to market shifts. ML addresses these challenges by enabling systematic analysis and pattern recognition at scale. It is, therefore, increasingly applied to tasks such as property valuation, market analysis, risk assessment, and personalized recommendations. ML tools extract valuable insights from vast amounts of data, enhancing decision-making across the sector. In the context of real estate, this includes analyzing historical sales data, property characteristics, market trends, and other relevant information to extract actionable insights that would otherwise be difficult or impossible to identify manually.

Both large corporations and smaller enterprises are focusing on consumer data, with interests ranging from locations and purchasing habits to overall satisfaction levels. Given the plethora of available data, proficiency in data mining techniques can yield valuable models that assist investors, sales agents, and prospective buyers alike**.**

2 Motivation

The motivation behind this research arises from the intersection of ML, real estate analytics, and financial risk assessment. Historically, the real estate industry has relied on traditional valuation methods that often fail to capture real-time market dynamics, resulting in inefficiencies and poor risk assessment during economic downturns. These inefficiencies can lead to overvalued assets, mispriced investments, delayed responses to market corrections, and increased exposure to financial losses. Without timely data-driven insights, stakeholders may overlook early warning signals of market stress, contributing to reduced portfolio performance or systemic vulnerabilities. This study addresses these challenges by applying advanced ML techniques to improve price prediction accuracy.

From an information systems perspective, this research tackles a core challenge: transforming fragmented, heterogeneous data into reliable, actionable intelligence for real estate decision-making. Specifically, it involves modeling the complex, non-linear relationships between property prices and external economic indicators such as inflation, unemployment, and interest rates. This challenge is not only an information systems concern, involving the design of data-driven tools to support decision-making under uncertainty. It is also a computer science problem, requiring advanced modeling techniques that can learn from high-dimensional, interdependent datasets. Traditional statistical methods, such as Linear Regression, struggle to integrate these multifaceted features effectively. By leveraging advanced ML approaches, including Gradient Boosting models (e.g., CatBoost), this study demonstrates how predictive analytics can support more adaptive, context-aware information systems in the real estate sector.

Additionally, the study was motivated by the increasing need for data-driven decision-making in real estate investments. As markets grow more volatile and are influenced by complex macroeconomic forces, relying solely on experience or heuristics is no longer sufficient. Investors, policymakers, and financial institutions require timely, objective insights to evaluate opportunities, mitigate risks, and adapt to shifting conditions. Accurately predicting property prices has, therefore, become essential for optimizing asset allocation, avoiding overexposure, and ensuring long-term portfolio stability.

Although this study focuses on property data from the U.S. market, it serves as a proof of concept for how advanced predictive models can be applied to other contexts to support more informed, evidence-based real estate decision-making. This study aims to apply ML to help buyers, sellers, and companies make more informed decisions, optimizing property pricing, identifying undervalued assets, and forecasting market downturns.

3 Scope

The study uses ML techniques to enhance decision-making processes in the real estate sector, specifically focusing on supervised learning-based pricing models. These models are designed to predict property Sale Amounts by analyzing internal factors (such as Assessed Value and Property Type) and external influences (such as market trends and ecomic factors like inflation). The predictive approaches evaluated in the study include a baseline Linear Regression model using internal property attributes, an enhanced regression model that incorporates external economic factors, and CatBoost, an advanced Gradient Boosting algorithm. Using a dataset of real estate transactions, the objective is to develop models that provide accurate and dynamic valuations, enabling more informed pricing decisions for investors, buyers, and sellers. This study concentrates on using ML to predict property prices. It does not cover other related areas, such as real estate risk scoring or macroeconomic crisis prediction, that may also benefit from utilizing ML techniques.

4 Research Background

Integrating ML and big data analytics has transformed the real estate industry, enhancing predictive capabilities, optimizing operations, and improving customer experiences. According to Mally (2023), big data analytics enables real estate professionals to forecast market fluctuations precisely. By analyzing extensive datasets, including historical price trends, economic indicators, and consumer behavior, ML tools surpass traditional methods reliant on manual assessments and a small set of comparable property sales selected by human appraisers. Mally (2023) noted that Automated Valuation Models use data like recent sales and neighborhood metrics to generate instant, objective valuations, playing a pivotal role in lending, investment evaluation, and taxation.

Choy and Ho (2023) emphasize ML's ability to accurately predict property values by incorporating diverse factors such as demographics, location, and property features. These algorithms support investment decisions by highlighting properties likely to appreciate or yield high rental incomes. Additionally, Geographic Information Systems combined with ML can uncover spatial patterns in property values, infrastructure development, and neighborhood dynamics that might remain undetected through traditional analysis. These insights help identify emerging investment zones, undervalued areas, or locations at risk of price stagnation.

ML also enhances property management operations, optimizing tasks like lease renewals and rent collection, as described by Choy and Ho (2023). Moreover, it aids in detecting fraudulent activities, ensuring transaction security and reliability. In customer engagement, real estate platforms increasingly rely on ML-based recommendation algorithms to personalize property searches. These algorithms analyze user behavior, preferences, and search histories to offer tailored suggestions, streamlining the buying experience.

From a technical perspective, Choy and Ho (2023) provide a detailed comparison of ML models used in real estate research, demonstrating that models like Extra Trees, k-Nearest Neighbours, and Random Forest significantly outperform traditional methods like Ordinary Least Squares in predictive accuracy and error minimization. Metrics such as mean squared error, root mean squared error, and mean absolute percentage error illustrate the superiority of ML models over traditional statistical techniques.

Such advancements have profound implications for sustainability in real estate. Accurate price signals generated by ML can guide governments in promoting energy efficiency and sustainable development. Buyers also benefit from these tools by avoiding overpriced properties and reducing waste in property development, as noted by Choy and Ho (2023). While the benefits of ML in real estate are evident, both Mally (2023) and Choy and Ho (2023) underscore the ethical challenges associated with its adoption.

4.1 Price Prediction in Real Estate

The application of ML in house price prediction has significantly advanced real estate analytics, providing data-driven insights for buyers, realtors, and investors. ML models evaluate various factors affecting house prices, such as property attributes, location, and amenities, with central location properties generally commanding higher prices. Studies by Sharma et al. (2023), Sangani et al. (2017), and Truong et al. (2020) emphasize how different algorithms handle these features and compare their performance across predictive tasks.

Different ML algorithms are foundational for real estate price predictions. Linear Regression serves as a baseline but struggles with complex datasets (Sharma et al., 2023; Sangani et al., 2017). Gradient Boosting Regressor and its variants, like XGBoost and Histogram Gradient Boosting Regressor, show superiority over Linear Regression due to their accuracy and precision (Sharma et al., 2023; Sangani et al., 2017). Random Forest and newer techniques, like LightGBM and hybrid models, enhance prediction accuracy, though Random Forest may be overfitting on test data [6]. Hybrid models combining Random Forest, XGBoost, and LightGBM achieved robust performance, balancing bias and variance effectively. Stacked generalization, combining predictions from multiple models, emerged as the most accurate approach (Truong et al., 2020). The robustness of these models depends significantly on data quality. Proper handling of missing data, dimensionality reduction, and feature encoding and augmentation enhance model performance (Sharma et al., 2023; Sangani et al., 2017; Truong et al., 2020).

Evaluating model performance using metrics like Mean Absolute Error and Root Mean Squared Error (RMSE) is crucial. Challenges remain, including data quality, algorithm complexity, and overfitting. However, by addressing these issues and employing novel methodologies, ML models can potentially revolutionize property valuation and decision-making in the housing market (Sharma et al., 2023; Truong et al., 2020). ML models can automate valuations at scale, adapt to changing market conditions in real-time, and uncover complex patterns that traditional methods often miss.

5 Methodology

5.1 Data Collection and Analysis

The dataset “Real Estate Sales 2001-2021”, derived from Data.gov (2022), underpins this study. It contains comprehensive real estate sales data for the state of Connecticut, covering a twenty-year period. Maintained by the Office of Policy and Management, this dataset includes all real estate transactions with sale amounts of $2,000 or more, collected under Connecticut General Statutes. This dataset was selected due to its size, coverage, and long historical range, which enable the exploration of both short- and long-term price dynamics. While it does not include granular property-level details such as square footage, number of rooms, or amenities, it provides consistent pricing and location data across a large and diverse region. At the time of selection, alternative datasets that included richer property features were either too small in scope or restricted to narrowly defined geographical areas, limiting their usefulness for generalizable model development. One limitation of the dataset is its regional scope, as it exclusively includes transactions from Connecticut. While this ensures internal consistency and a rich historical record, it may introduce regional biases that limit generalizability to other markets. However, the methodology is designed to be transferable and can be adapted for similar datasets in other geographic regions.

The dataset comprises 997,213 entries. Table 1 provides an overview of each feature used from the Real Estate Sales 2001-2002 dataset, including the number of missing values, the number of valid entries, and a brief description of what each feature represents.

Table 1. Dataset Features

Attribute Name	Null Count	Non-Null Count	Description
Serial Number	0	997213	Unique identifier for each sale record.
List Year	0	997213	The year in which the sale was recorded.
Date Recorded	2	997211	The date the sale was recorded.
Town	0	997213	The town where the property is located.
Address	51	997162	The specific address of the property.
Assessed Value	0	997213	The valuation of a property determined by a government assessor for taxation purposes.
Sale Amount	0	997213	The recorded transaction price at which the property was sold.
Sales Ratio	0	997213	The ratio of Sale Amount to Assessed Value.
Property Type	382446	614767	Categorization of the property (residential, apartment, commercial, industrial, or vacant land).
Residential Type	388309	608904	Further classification of residential properties.
Non-Use Code	707532	289681	Code indicating non-standard use of the property.
Assessor Remarks	847359	149854	Additional remarks provided by the assessor.
OPM remarks	987279	9934	Remarks provided by the Office of Policy and Management.
Location	799516	197697	General geographic coordinates (such as latitude and longitude) used to locate the property on a map.

To enhance model robustness, datasets of external economic factors were incorporated, accounting for various macroeconomic influences on the real estate market. These datasets provide insights into market conditions, inflation, interest rates, and overall economic stability (U.S. Bureau of Labor Statistics, 2024; World Bank, 2024; Federal Reserve Bank of St. Louis, 2024; U.S. Census Bureau & U.S. Department of Housing and Urban Development, 2024).

5.2 Data Preparation and Pre-Processing

The data preparation phase involved a series of crucial steps to ensure the dataset's readiness for analysis. Python libraries were employed for pre-processing tasks, with Pandas (a data analysis and manipulation tool built on top of Python) used for data manipulation and analysis. These steps align with widely adopted practices in data science and predictive modeling, as outlined in works such as Han et al. (2012). Initially, the dataset was cleaned to address missing values, convert categorical attributes into numerical formats, and identify and remove outliers.

A key transformation during pre-processing was the decision to vectorize Property Type instead of Town. This approach was adopted due to high cardinality in the Town attribute, which would have led to unnecessarily long and sparse vector representations. In contrast, Property Type had a limited set of categories, allowing for a more compact and meaningful representation.

To enhance the dataset, feature engineering was applied by introducing a new attribute, Listing Age, which represents the difference between the current year and the year of listing. This feature provided insights into how long properties remained on the market and improved predictive model performance.

Columns with minimal relevance or excessive missing values were removed, reducing noise and improving overall model performance. Missing values in numerical attributes were replaced with the mean, while those in categorical attributes were filled with the most frequent value. Further consistency checks ensured all recorded values fell within reasonable ranges, removing negative values and flagging sale amounts deviating significantly from historical averages.

Outlier detection, focused on Sales Ratios, was critical in maintaining data integrity and model reliability. The primary objective was developing accurate ML models for price prediction, requiring the exclusion of transactions with unrealistic Sales Ratios and ensuring that the models were trained on market-driven sales rather than anomaly transactions.

5.3 Price Prediction Models

This section details the methodology for developing price prediction models using ML techniques, specifically Linear Regression and more advanced models. These models estimate property sale amounts based on various internal and external factors.

5.3.1 Linear Regression Model

Linear Regression serves as the foundation for the price prediction models, aiming to predict property prices based on key features: Assessed Value, List Year, and Property Type. This model assumes a linear relationship between these predictors and the target variable, Sale Amount, enabling price estimation based on historical data.

The general form of multiple Linear Regression, as presented by Han et al. (2012), is expressed as:

$$y = \beta_{(0)} + \beta_1 x_1 + \beta_2 x_2 + + \beta_n x_n + \epsilon$$

where:

- y represents the predicted Sale Amount.
- x_1 , x_2 , ... , x_n are the input features (e.g., Assessed Value, List Year, Property Type).
- β_0 *is* the intercept, representing the expected value of y when all predictors are zero.
- β_1 β_2 , ... , β_n are the coefficients representing the impact of each feature on the Sale Amount.
- ϵ is the error term, accounting for variability not captured by the model.

The independent variables (x) include Assessed Value, List Year, and Property Type, while the dependent variable (y) corresponds to the Sale Amount. Property Type, a categorical variable, was one-hot encoded to create binary dummy variables for the regression model.

The Linear Regression model used in this study is implemented using the Scikit-learn Python library, as described by Pedregosa et al. (2011). To divide the dataset into training and testing sets, the train_test_split function from Scikit-learn was utilized, with 80% of the data allocated for training the model and 20% reserved for testing its performance. The Linear Regression Model was instantiated and trained on the training data using the Fit Method. To ensure reliable results, the model was evaluated using Cross-Validation, a technique that tests the model on multiple subsets of the data. This approach helps verify that the model's performance is consistent and not overly dependent on one slice of the dataset. After training, predictions were generated for both the training and testing sets. The model's performance was evaluated using Mean Squared Error (MSE), R^2 scores, and accuracy, assessing the model's prediction strength across different datasets.

5.3.2 CatBoost

CatBoost is an ML algorithm built on the Gradient Boosting Ensemble Method. Ensemble methods combine the outputs of many simpler models to produce more accurate and robust predictions than any single model alone. In this case, the simpler models are Decision Trees, which split data based on feature values to make predictions. Each individual tree is known as a weak learner because it performs only slightly better than random guessing, but when combined in sequence, their errors are gradually corrected, improving overall performance. Unlike traditional Gradient Boosting algorithms, CatBoost introduces innovations such as Ordered Boosting and Permutation-Based Processing, which help reduce prediction bias and improve learning efficiency, particularly when working with categorical data. While XGBoost and LightGBM are widely used in similar prediction tasks, CatBoost was selected due to its native support for categorical features, reduced need for extensive pre-processing, and competitive performance in recent literature. Furthermore, existing studies have already explored real estate price prediction using XGBoost. At the same time, CatBoost remains underutilized in this domain, providing an opportunity to contribute novel insights using a state-of-the-art method.

Gradient Boosting builds a model incrementally by adding one weak learner at a time. In each new iteration, a new tree, $h_t(x)$ is trained to correct the residual errors from the previous iteration. As introduced by Prokhorenkova et al. (2018), in the context of CatBoost, the model is updated using the following formula:

$$F_t(x) = F_{(t-1)}(x) + \gamma h_t(x)$$

where:

- $F_t(x)$ is the predicted function at iteration t.
- $h_t(x)$is the newly trained Decision Tree model.
- γ is the learning rate, controlling the contribution of each new tree to the final prediction.

To improve the model's predictive accuracy, key settings within the CatBoost algorithm, such as learning rate, tree depth, and regularization strength, were systematically adjusted. These tuning efforts aimed to strike a balance between performance and training efficiency, ensuring the model did not overfit the data while maintaining general reliability.

CatBoost was chosen to improve upon the limitations of Linear Regression, particularly its ability to handle categorical features natively. In earlier stages of pre-processing, Town was excluded as a feature due to its high cardinality, which would have required encoding it into hundreds of sparse binary columns using traditional one-hot encoding. This would have introduced unnecessary complexity and reduced model efficiency. However, CatBoost can process categorical features directly without explicit encoding, which allows the Town feature to be reintegrated into the model in a more efficient and meaningful way.

This model demonstrated superior predictive accuracy compared to Linear Regression, capturing complex feature interactions and enhancing reliability.

6 Results and Analysis

The CatBoost model outperformed other predictive models, effectively capturing complex relationships between features. The Assessed Value feature emerged as the strongest predictor, confirming its role as a primary determinant of Sale Amount. The Town feature also emerged as a significant feature, underscoring the importance of location in real estate valuation.

Feature importance analysis highlighted the need for careful feature selection and encoding choices, influencing the performance of ML models. The reintegration of Town was a valuable adjustment, improving predictive accuracy and highlighting that location remains a fundamental driver of real estate pricing. The Pearson correlation coefficient between Assessed Value, List Year, Town, Property Type, and Sale Amount was calculated and is presented in Table 2.

Table 2. Pearson Correlation Between Key Features and Sale Amount

Feature	Pearson Correlation with Sale Amount
Assessed Value	0.89 (Strong Positive)
List Year	0.41 (Moderate Positive)
Town	0.35 (Moderate Positive)
Property Type	0.28 (Weak Positive)

The strongest correlation was observed between Assessed Value and Sale Amount, which aligns with expectations, as assessed property values generally serve as a basis for market price estimations. List Year and Town also exhibited moderate correlations, suggesting that newer listings and geographical location played roles in determining Sale Amount. Property Type showed a weaker correlation, indicating that while it influences pricing, other external factors likely introduce more variation.

While correlation analysis provides a general understanding of linear relationships, ML models can capture complex non-linear interactions between features. CatBoost assigned importance scores to each feature based on how frequently and effectively they contributed to reducing the model's error. The importance scores of different features, as these were generated by the trained CatBoost model using test data, are shown in Table 3.

Table 3 CatBoost Feature Importance Scores

Feature	Importance Score
Assessed Value	52.7%
Town	18.2%
List Year	11.6%
Property Type	9.5%
External Factors	8%
Assessed Value	52.7%

The comparison between Linear Regression and CatBoost demonstrates distinct approaches to feature engineering. Linear Regression identified Assessed Value as the most influential predictor due to its high correlation with Sale Amount. Still, it overlooked the significance of the Town feature, which was excluded due to high cardinality. CatBoost addressed this by dynamically adjusting feature importance, revealing Town as the second most significant factor in pricing. This analysis revealed several insights. Assessed Value consistently predicted prices best. Town was critical for price differentiation. Property type was less influential than expected. External factors, like inflation, unemployment, and interest rates, had limited individual impact but were meaningful in combination. This underscored the importance of selecting appropriate features and encoding techniques, as these choices significantly impacted model performance. Reintroducing Town in the CatBoost model improved accuracy and reaffirmed the importance of location in real estate pricing.

6.1 Error Analysis and Model Comparison

The error analysis and model comparison highlighted the progressive improvement and refinement achieved through enhanced modeling techniques and feature engineering.

As shown in Figure 1, the Initial Regression Model, constrained by limited features, demonstrated poor predictive power with an R^2 of 0.31 and high RMSE, reflecting significant deviations in predictions. Introducing pre-processing and additional features moderately enhanced performance, with an R^2 of 0.56 and reduced RMSE, yet the Improved Regression Model still lacked optimal reliability. The CatBoost model, leveraging advanced algorithms and feature interactions, emerged as the most robust, achieving an R^2 of 0.78 and a low RMSE.

These results underscored the importance of iterative model refinement and the adoption of advanced ML techniques for addressing complex predictive tasks. This demonstrated the potential of modern ensemble methods like CatBoost to deliver highly reliable predictions compared to traditional regression models. While the CatBoost model outperformed other approaches in terms of R^2 and RMSE, further evaluation using cross-validation techniques confirmed the consistency of its performance. In cross-validation, the training data is split into multiple subsets and the model is trained and validated across these different partitions to ensure generalizability. The model's predictions remained stable across subsets, with minimal variance, and achieved a Mean Absolute Error of 9.74% on the test set. These results suggest that the model's improvements were not due to overfitting but reflect predictive strength.

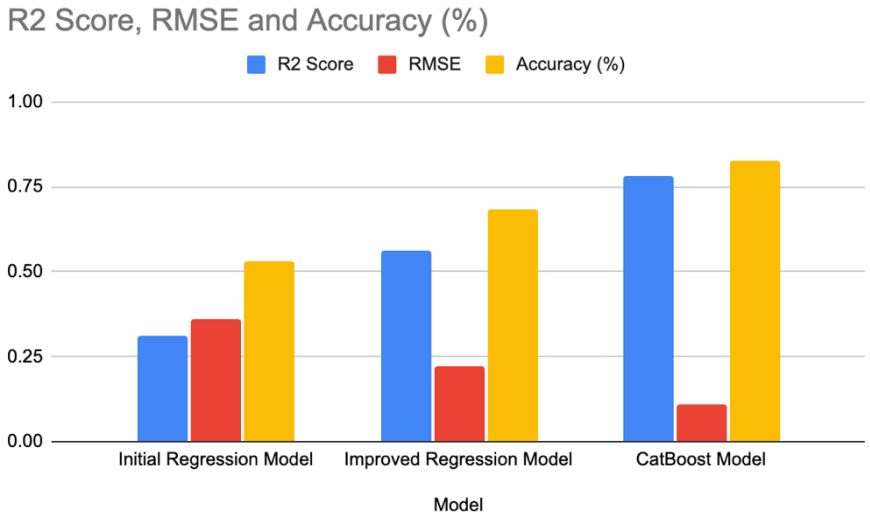

Figure 1. Comparison of model performance across Linear Regression, Improved Regression, and CatBoost using R^2 and RMSE metrics

The analysis underscored the enhanced accuracy achieved through ML-driven methods, particularly when advanced models like CatBoost are utilized.

Predictions were compared across the three models using two sample properties to evaluate model accuracy. A publicly available prediction by Zillow (2024), referred to as Zestimate, and an associated Prediction Range were used as a benchmark. As shown in Table 4, the Initial Regression Model, which used only internal features (Assessed Value and List Year), produced highly generalized results. The Improved Regression Model incorporated additional features but lacked nuance in feature interaction. The CatBoost model, which combined internal and external features within a Gradient Boosting framework, delivered the most reliable predictions, closely aligning with Zillow's Prediction Range, and a deviation of 3.21% from Zillow's Prediction, Zestimate.

Table 4. Predicted property prices using the three models and compared with Zillow's Zestimate and prediction range

Property Name	Initial Model's Prediction	Improved Regression Model's Prediction	CatBoost's Prediction	Zillow's Prediction (Zestimate)	Zillow's Prediction Range
158 4th Ave, Milford, CT 06460	$725,000	$860,000	$912,000	$907,500	$844,000 - $989,000
9 Hinchley Wood, Farmington, CT 06032	$1,250,080	$1,480,000	$1,600,000	$1,526,300	$1.37M - $1.71M

The CatBoost model's strong performance align with prior research on the effectiveness of Gradient Boosting for non-linear data. This result highlights the value of ensemble methods in improving the accuracy of real estate price prediction.

7 Discussion

This study contributes to the information systems literature by demonstrating the application of ML as a decision-support mechanism in real estate valuation. It shows that supervised learning methods, particularly ensemble techniques like CatBoost, can significantly improve pricing accuracy over traditional models by capturing non-linear interactions and integrating heterogeneous data sources.

The real-world value of this research lies in its practical applicability to a range of stakeholders in the property market. Real estate investors can use predictive models like CatBoost to identify mispriced properties, anticipate returns, and minimize valuation risk. Developers can incorporate model outputs to navigate pricing decisions during project planning, while real estate agents can benchmark property prices during

negotiations. Notably, the close alignment between the CatBoost model's predictions and Zillow's estimates reinforces the practical validity and potential for adopting the CatBoost model in real estate price projections.

From a systems perspective, this research shows how information systems can use ML to turn fragmented sales data into actionable real estate insights. Although trained on U.S. data, the methodology demonstrates a transferable framework that could be applied in similar data-limited contexts, serving as a proof of concept for broader deployment.

8 Limitations and Future Work

While the results demonstrate the effectiveness of ML in predicting property prices, several limitations must be acknowledged. First, data quality posed a recurring challenge. The dataset spans two decades of real estate transactions, during which inconsistencies in feature availability and data collection standards were observed. Older records often lacked complete information, and external economic indicators required significant pre-processing to align them with transaction-level data. The exclusion of certain features due to high cardinality (such as Town in early models) also limited initial model performance. Moreover, while the models performed well on historical data, their generalizability to other geographic regions or rapidly evolving market conditions has not yet been tested. Finally, ethical concerns around data-driven valuation models, such as potential reinforcement of market biases through algorithmic predictions, should be carefully considered in their future development and deployment.

Future work should focus on enriching the models with additional property-level attributes to improve prediction accuracy. Incorporating geospatial data into the model could enable localized trend analysis and interactive mapping of predicted prices and market risks. Additionally, connecting the model to real-time data streams, such as live property listings and macroeconomic feeds, would support continuous model updates, improve responsiveness to market shifts, and enable the development of intelligent alert systems for investors and policymakers. These enhancements would pave the way for a real-time, AI-powered real estate analytics platform capable of supporting dynamic, data-driven decisions at scale.

9 Conclusions

This study demonstrates that ML offers a viable and effective approach to property price prediction in real estate analytics. By incorporating both internal and external factors and leveraging ensemble-based modeling techniques, the research shows how predictive accuracy can be substantially improved over traditional regression models.

The key contribution lies in applying ML within an information systems context to create data-driven tools that support real-world decision-making. The models developed here serve as a proof of concept for how predictive analytics can enhance transparency, reduce valuation uncertainty, and operationalize large-scale historical datasets in the property market.

As data becomes increasingly central to investment planning, pricing, and development decisions in the real estate sector, this work underscores the growing role of intelligent systems in shaping evidence-based decisions for investors, developers, and public institutions alike.

Declaration of AI

While preparing this work, the authors used ChatGPT (OpenAI) to produce an initial summary of a longer thesis document. The prompt instructed the model to preserve the original language and style. The generated output was reviewed, verified for accuracy, and substantially edited by the authors to ensure factual correctness and alignment with both scholarly standards and the original thesis from which the content was derived.

References

Arthur, M. (2022). How accurate is my Zestimate, and can I influence it? Retrieved from Zillow: https://www.zillow.com/learn/influencing-your-zestimate/

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

Data.ct.gov. (2022). Real estate sales 2001-2022 GL [Dataset].

Federal Reserve Bank of St. Louis. (2024). 10-year treasury constant maturity minus 2-year treasury constant maturity [T10Y2Y].

Han, J. K. (2012). Data Mining: Concepts and techniques. Elsevier.

Mally, P. (2023). Data and algorithms: Reviewing the role of machine learning in the real estate sector. International Journal of Computer Trends and Technology, 71(11), 55-64.

Pedregosa, F. V. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Prokhorenkova, L. G. (2018). CatBoost: Unbiased boosting with categorical features. arXiv preprint arXiv:1810.11363.

Sangani, D. E. (2017). Predicting Zillow estimation error using linear regression and gradient boosting. IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS 2017).

Sharma, S. A. (2023). House price prediction using machine learning algorithm. 6th Intenational Conference on Computing Methodologies and Communication (ICCMC 2022), (pp. 982-986).

Truong, Q. N. (2020). Housing price prediciton via improved machine learning techniques. Procedia Computer Science, 174, 433-442.

U.S. Bureau of Labor Statistics. (2024). Unemployment rate [UNRATE]. Federal Reserve Bank of St. Louis.

U.S. Census Bureau & U.S. Department of Housing and Urban Development. (2024). Average sales price of houses sold for the United States [ASPUS].

U.S. Census Bureau & U.S. Department of Housing and Urban Development. (2024). Median sales price of houses sold for the United States [MSPUS].

U.S. Census Bureau. (2024). Median household income in the United States [MEHOINUSA646N].

World Bank. (2024). Inflation, consumer prices for the United States [FPCPITOTLZGUSA]. Federal Reserve Bank of St. Louis.