Can machine learning improve state-level house price forecasts compared with traditional forecasting methods?

Nathan Vancauwenbergh
natecoldmountain@gmail.com

ABSTRACT

This paper tests whether machine learning is able to forecast real estate better than human judgment. A state-level panel dataset from January 2000 until early 2026 will be utilized where the dependent variable of Zillow Home Value Index will be merged with price momentum, volatility, building permits, income, unemployment rate, growth of CPI, state fixed effects, and seasonality.

Methodology includes three models – no change forecast, OLS regression and Random Forest machine learning algorithm. According to findings, Random Forest outperformed no change forecast in all 50 states and Washington DC, but did not outperform OLS regression with state fixed effects in many cases. In addition, OLS Regression demonstrated lower Mean Absolute Error whereas Random Forest had lower Root Mean Squared Error, which means that machine learning helped to minimize bigger errors.

INTRODUCTION

In this project, real estate data and economic variables are analyzed in order to examine the ability of AI models to make forecasts. Specifically, the focus will be put on house prices, real estate returns, and risk factors. Agentic coding is applied in the development and testing of models, including the random forest. Unlike other papers examining AI, this project does not aim to answer the question "is it possible for AI to predict real estate prices". It aims to understand whether it is possible for the AI to uncover useful patterns in real estate investments; whether it performs better than naive algorithms (e.g., considering the previous day prices) and whether it takes risks which the human investor usually avoids. This makes the project important since real estate may bring profits as well as large losses. Real estate assets become more valuable and can bring profits in the form of rent, but they are rather expensive and illiquid. Moreover, there are numerous risk factors which should be considered when making investment decisions.

Real estate is different from many other investments because it is relatively illiquid, meaning a property owner cannot quickly turn a property into cash. When an investment underperforms, a stock can often be sold very fast, whereas a property can take months to sell. There can also be taxes, legal costs, and other fees. This makes real estate risky because a bad investment decision can keep an investor stuck for a long time.

Real estate returns are also hard to predict because many factors affect them simultaneously. A property can become more valuable when the area becomes more popular, rents increase, the economy is doing well, or the building itself is improved. However, a property can also lose value because of higher interest rates, lower demand, unemployment, excess supply, regulation, climate risk, or a weaker economy. Investors must consider both profit and risk. Usually, higher risk should mean higher expected return, but not every risky investment is worth taking.

A simple way to predict house prices is to use past prices. In economics and statistics, this is similar to an autoregressive approach, where future values are predicted using earlier values. For example, if a house was worth a certain amount yesterday, it may be worth a similar amount today. This can be useful because real estate prices usually do not change suddenly from one day to the next. However, this approach has limits because it does not explain why prices change. If interest rates rise, demand falls, or the local economy weakens, past prices may no longer be a reliable guide. This is why this study also considers other factors that may affect real estate returns, rather than relying only on previous prices.

Investment companies that invest in various kinds of real estates like residential, office space, hotels, logistics, and commercial real estates are Blackstone, Brookfield Asset Management, Starwood Capital Group, KKR, Apollo Global Management, and Tishman Speyer. These companies conduct risk assessment before making an investment decision. They consider issues such as if there is a growth in demand, if the rents are going up, if the location is improving, if there is oversupply, and if it can be resold.

Such ventures seek to invest in projects where potential gains outweigh potential risks. For instance, an organization will be more likely to invest in properties used for housing when there is a high demand for rentals. Other entities will invest in logistics where warehouse spaces are needed for online retail businesses. Others will engage in investments involving distressed properties despite them being risky and potentially profitable.

AI is therefore useful to explore in this context because whereas humans utilize their experience and intuition in local knowledge and judgement, there are factors that may lead to mistakes. Humans might be overconfident in their choices or may make decisions based on current trends.

AI works differently. Instead of using instinct, AI looks for patterns in data. A machine learning model can compare many factors at the same time. One model I will use is random forest. Random forest builds many decision trees and combines them to make a prediction. This can help because real estate is affected by many variables at once. For example, random forest may help test whether house prices, income, unemployment, housing supply, or risk factors are important for predicting real estate returns.

Although this research suggests that AI can be useful in real estate forecasting, less attention has been given to when AI may fail, how its decisions compare with human judgement, and what types of risks AI may take that human investors would normally avoid. This directly supports the central research question: can machine learning outperform human judgement in real estate forecasting, and under what market conditions does its advantage become less reliable?

DATA

To evaluate whether machine learning models can improve state-level house price forecasts relative to traditional econometric frameworks, this study constructs a multi-dimensional panel dataset covering the period from January 2000 to early 2026. The dataset is used to test whether Random Forest can predict house prices more accurately than simpler models, and to compare how different variables, such as past price changes, income, unemployment, and housing supply, affect the forecasts.

Real estate markets are influenced by more than historical prices alone. House price performance is shaped by local demand, supply conditions, macroeconomic variables, spatial differences, and historical price momentum. Therefore, relying only on lagged house prices would provide an incomplete view of market risk. The dataset was designed to record these different dimensions by combining a target house price index, momentum-based features, supply and demand indicators, macroeconomic controls, and categorical controls for geography and seasonality.

The dataset's 26-year span records several distinct housing market regimes, including the early-2000s appreciation cycle, the 2008 Global Financial Crisis, the post-crisis correction and recovery, and the post-COVID inflationary period. This allows the models to be tested across both stable and unstable market environments. As a result, the dataset can be used to assess whether machine learning models maintain predictive strength across different regimes, or whether their performance deteriorates during structural shocks and less expected market conditions.

The graph provides an illustrative example of the housing market cycles captured in the broader dataset, using selected New York City home values. It shows that values rose strongly before the 2008 financial crisis, declined by 2011, and then recovered over the following years. This shows that real estate prices can rise, fall, and recover over time, which makes predicting real estate returns difficult.

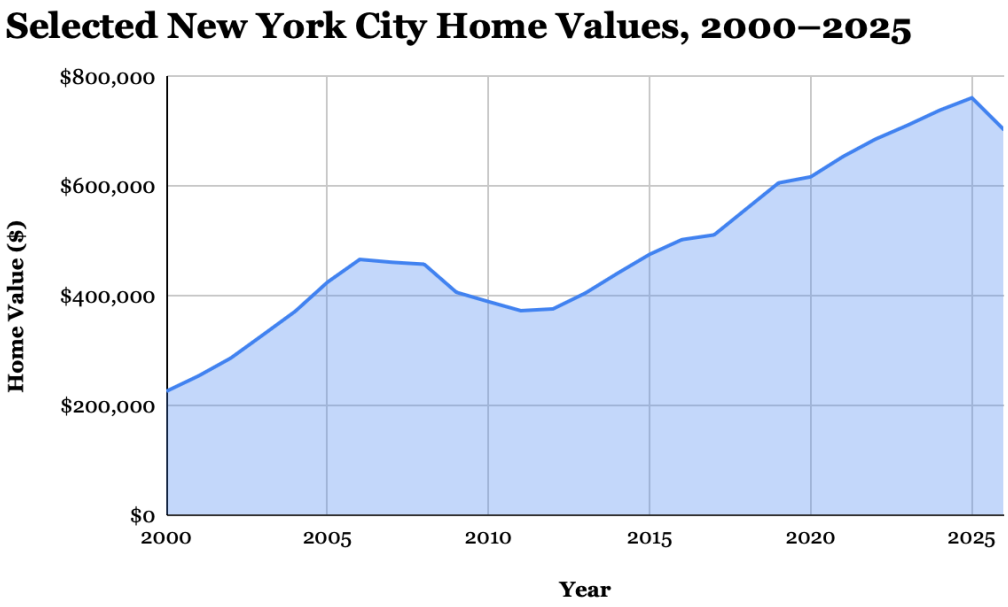


Figure 1. Selected New York City Home Values, 2000–2025.
Figure 1, shows the house prices illustrative for New York City, however data can be available to all other states. Showing that selected New York City home values rose before 2008, declined after the financial crisis, and later recovered. It is an illustrative example; the forecasting models use state-level data for all 50 states and Washington, D.C.

The central target variable is the Zillow Home Value Index (ZHVI). ZHVI is used as the main measure of residential real estate value across state-level markets. It is useful because it provides a more stable measure of housing values than raw transaction prices, which can change depending on the type, quality, and location of homes sold in a given month. This makes ZHVI suitable for comparing housing trends across the 50 U.S. states and Washington, D.C.

To capture past price dynamics, the ZHVI series was transformed into multiple variables, including 1-month, 3-month, 6-month, and 12-month trailing returns. A historical price volatility measure was also calculated using the standard deviation of the index. These features allow the models to test whether short-term and medium-term price movements contain useful predictive information. They also allow the study to compare machine learning models with simpler approaches, such as a naive autoregressive benchmark that assumes future prices will closely follow recent past prices.

The dataset also includes supply-side and demand-side risk variables. New private housing building permits are included as a supply-side indicator. Building permits are relevant because they provide information about expected future housing supply. A decline in permits may indicate future supply constraints, while an increase in permits may signal potential oversupply and weaker future price growth.

On the demand side, the dataset includes median household income and unemployment rates. These variables are used to measure local economic strength and household purchasing capacity. Rising incomes and low unemployment can support housing demand, while increases in unemployment can weaken demand by reducing credit capacity and household confidence.

To account for broader macroeconomic conditions, the dataset includes Consumer Price Index growth. CPI growth is relevant because inflation affects construction costs, household purchasing power, and the broader financing environment. These factors are important in real estate markets, where valuations are highly sensitive to affordability, interest-rate conditions, and expectations about future economic stability.

The panel also includes state fixed effects and month seasonality controls. State fixed effects account for persistent cross-sectional differences between states, including tax structures, regulatory frameworks, geographic constraints, and other slow-moving local characteristics. Month seasonality controls account for recurring calendar patterns in housing markets, such as seasonal differences in buying activity, listing behavior, and transaction timing.

By combining historical price data with supply, demand, macroeconomic, spatial, and seasonal variables, the study is able to examine not only whether machine learning models produce accurate forecasts, but also which types of information appear most relevant to their predictions. This allows the paper to consider whether models such as Random Forest and XGBoost respond mainly to recent price trends or whether they also capture wider economic and market conditions. This directly supports the central research question: can machine learning outperform human judgement in real estate forecasting, and under what market conditions does its advantage become less reliable?

Variable	Frequency	First Observation	Min	N. of observations	Mean	Max	Source
Zillow Home Value Index	Monthly	Jan 2000–Mar 2026	$73,950.95	15,837	$231,279.78	$854,559.86	Zillow
1-month ZHVI return	Monthly	Feb 2000–Mar 2026	−3.41%	15,781	0.34%	4.01%	Calculated from Zillow

12-month ZHVI return	Monthly	Jan 2001–Mar 2026	−30.55 %	15,220	4.37%	40.42%	Calculated from Zillow
ZHVI volatility	Monthly	Jan 2001–Mar 2026	0.014%	15,165	0.232%	1.674%	Calculated from Zillow
Building permits	Monthly	Jan 2015–Jan 2026	5	6,783	2,283.92	24,774.51	U.S. Census / FRED
Median household income	Annual	2015–2026	$27,213	7,488	$80,424.35	$109,707	Census / ACS / FRED
CPI growth	Monthly	Jan 2015–Mar 2026	234.747	7,581	272.856	330.293	BLS / FRED

Table 1. Summary statistics for the state-level panel dataset, January 2000–March 2026, except for median household income and CPI growth, which cover shorter periods. Because median household income is reported annually, each annual value was applied to all monthly observations within that year.

METHODOLOGY

The aim of this project is to analyze basic methods of forecasting against machine learning models to check whether models that use AI can find any patterns in price prediction of real estates. Large language models have been employed to help produce and refine Python scripts to do preprocessing, modeling, and comparison of results.

There are several ways to predict future house prices. The simplest method is to assume that the house price in the next period will be the same as the house price in the previous period. This is used as a no-change benchmark and can be written as:

$$y_t = y_{t-1}$$

In this equation, y_t is the house price at time *t*, and y_{t-1} is the house price in the previous period. This model is useful as a benchmark because house prices often move slowly over time. However, it also makes a strong assumption: it assumes that the change in house prices is zero.

This can be shown as: $y_t = y_{t-1}$ then: $\Delta y_t = y_t - y_{t-1} = 0$

Meaning that the benchmark predicts no increase or decrease in house prices. While this may work reasonably well during stable periods, it may fail when market conditions change, such as during periods of rising interest rates, inflation, changes in housing supply, or shifts in demand.

Since house prices may be affected by more than just the previous price, this project also uses regression models that include additional explanatory variables. A simple linear regression model can be written as:

$$y_t = \alpha + \beta X_t + \varepsilon_t$$

Here, α is the intercept, X_t represents factors that may explain house prices, β measures the relationship between those factors and house prices, and ε_t is the error term. In this project, the explanatory variables include previous house prices, price momentum, volatility, building permits, inflation, income, tax variables, state fixed effects, and month seasonality. One of the variables inside X_t can also be y_{t-1}, meaning the model can still use past house prices while also considering other factors.

In expanded form, this can be written as:

$$y_t = \alpha + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \ldots + \varepsilon_t$$

For example, $X_{1,t}$ could represent the previous month's house price, $X_{2,t}$ could represent inflation, and $X_{3,t}$ could represent building permits.

The project then compares these simpler models with a machine learning model, Random Forest. Random Forest works by building many decision trees and combining their predictions. This makes the model more stable than relying on one single decision tree and allows it to capture more complex patterns between variables. For example, the effect of building permits may be different in high-growth states compared with slower-growth states.

The data is segmented chronologically into training and test datasets. The observations that occur before January 2023 are included in the training dataset, while those that occur after January 2023 are included in the test dataset. The models are fitted solely based on the training dataset and are used to make one-month ahead predictions for the test period. The chronological segmentation makes sure that no future observations are used in the training process.

After the models make predictions, their forecasts are compared with the actual house prices. Model performance is evaluated using common accuracy measures such as Mean Absolute Error, Root Mean Squared Error, Mean Absolute Percentage Error, and R^2. These measures help show whether the machine learning model performs better than the simple benchmark and traditional regression models.

This means that the analysis is done to determine the variables the model takes as most significant. This is relevant in determining the answer to the research question because the research question asks not only whether the use of AI in the prediction of house prices is possible but also what kind of information and risks are taken into consideration by the use of AI.

LITERATURE REVIEW

There have been several studies concerning artificial intelligence and real-estate prediction. Tekouabou et al. (2024) show that machine learning is widely used in house-price prediction, real-estate valuation and decision-support systems. The authors demonstrate AI's ability to process large amounts of data and identify patterns in real-estate markets. However, they also identify limitations related to dataset size, data quality and a lack of explainability. This study is relevant to the present paper because the issue is not only whether AI can predict house prices, but also whether its predictions can be understood and explained.

Traditional real-estate valuation theory also demonstrates that human judgement is imperfect. According to Mooya (2016), property valuations may be influenced by uncertainty, real-estate bubbles, anchoring bias and client influence. Therefore, human judgement should not automatically be treated as a perfectly accurate benchmark when comparing human and machine-learning forecasts.

Real-estate returns are also affected by external economic factors. Ling and Naranjo (1997) find a relationship between commercial real-estate returns and macroeconomic risk factors, including interest

rates, inflation and broader economic conditions. Consequently, forecasting models should not rely only on previous prices. They should also incorporate supply, demand and other macroeconomic variables.

Market illiquidity is another important real-estate risk factor. Lin and Vandell (2007) explain that properties can be difficult to sell and that transaction processes may take a long time. As a result, historical transaction data may not accurately reflect current market conditions. Machine-learning systems may therefore overlook liquidity risks that are not visible in past data, whereas human investors may consider whether an asset can realistically be sold. Mukhamediev et al. (2022) further explain that machine-learning systems face limitations involving data quality, modelling problems and difficulty responding to sudden events. Regulatory changes, shifts in liquidity and other unexpected developments may not be captured in historical datasets.

Regarding the modeling technique used in this research, Wang (2021) examines the performance of OLS linear regression against Random Forest in house price prediction, using a dataset with 21,613 American properties and 20 housing features. It was found that Random Forest outperformed OLS in terms of forecasting accuracy, implying that machine learning techniques may prove efficient in case the connection between housing features and prices is complicated and nonlinear. However, Wang pays attention to the individual prices of the houses based on their housing and geographic features and does not consider the forecast of the dynamics of housing prices in general. Hence, there is a possibility to explore whether the superiority of Random Forest in forecasting compared to traditional statistical techniques is valid for the forecast of changes in the aggregate house prices.

This research will be devoted to the comparison of Random Forest and OLS, including state fixed effects and a no change benchmark, in the context of predicting changes in house prices in the states of America. Contrary to many previous studies devoted to the evaluation of individual houses' prices based on their characteristics, in this study the emphasis will be made on the forecasting of the state-level dynamics of housing prices based on the latest price changes and economic/housing market variables.

RESULTS

The results show that machine learning improved house price forecasting compared with the simplest benchmark, but it did not clearly outperform all traditional models. On the test set, the no-change benchmark had a Mean Absolute Error of about **$1,061** and a Root Mean Squared Error of about **$1,357**. OLS with state fixed effects produced the lowest Mean Absolute Error, at about **$379**, while the Random Forest model produced the lowest Root Mean Squared Error, at about **$582**. This means that Random Forest was useful, especially for reducing larger errors, but OLS remained highly competitive.

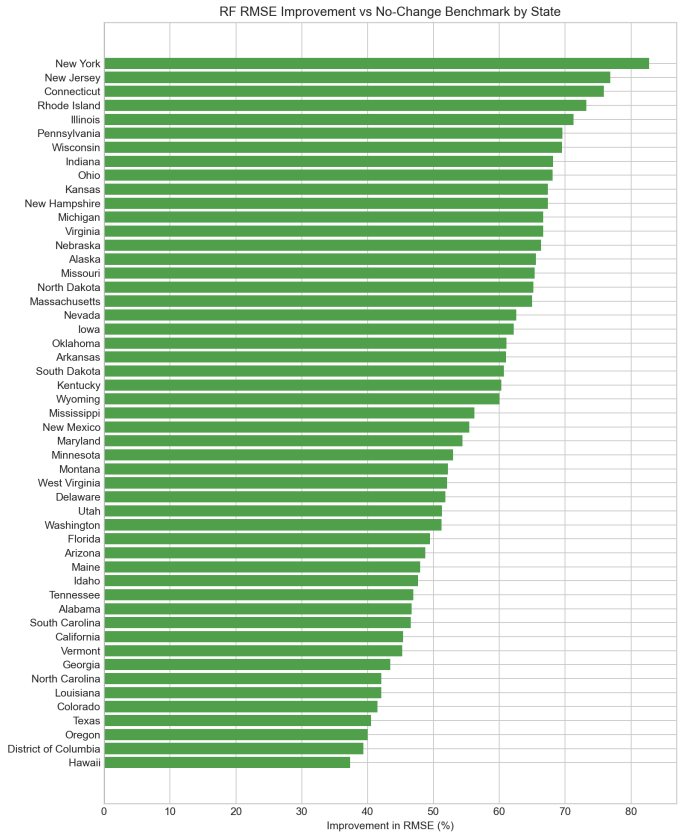


Figure 2. Random Forest RMSE Improvement Compared with the No-Change Benchmark by State
The positive values prove that the Random Forest improved prediction accuracy over the baseline method.

Figure 2 shows that Random Forest improved on the no-change benchmark in every state. This is important because it shows that the machine learning model added value compared with the simplest possible forecasting method. The improvement was especially strong in states such as New York, New Jersey, Connecticut, and Rhode Island, where recent price patterns helped the model reduce prediction error.

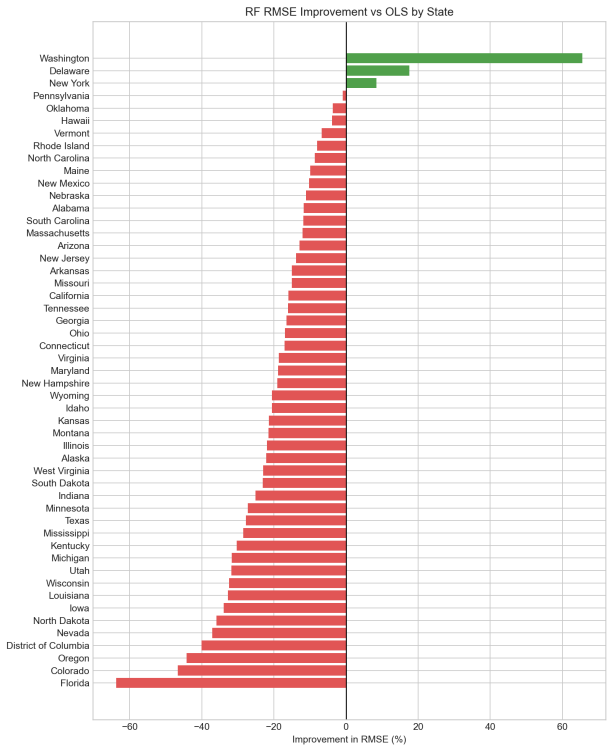


Figure 3. Random Forest RMSE Improvement Compared with OLS by State
The positive values are where Random Forest performed better than OLS, whereas negative values are where OLS performed better.

Figure 3 shows a more mixed result. Random Forest only outperformed OLS in a small number of states, including Washington, Delaware, and New York. In most states, the Random Forest model performed worse than OLS. This suggests that machine learning was better than the simple benchmark, but it did not consistently beat a well-designed traditional regression model with state fixed effects.

Some statelevel results show that model performance varied by region. For example, OLS with state fixed effects performed best in states such as Alabama, Arizona, and Arkansas, while Ridge Regression performed best in Alaska. This suggests that no single model worked best in every local housing market. Instead, the best model depended partly on the state being predicted. This is important because real estate markets are local: states can differ in price levels, supply conditions, income, population growth, and exposure to economic shocks.

LIMITATIONS AND FURTHER RESEARCH

This project has some limitations because real estate investing is complex and difficult to fully measure with data. The model can use variables such as house prices, interest rates, unemployment, income, housing supply, and risk factors, but it cannot include everything that a human investor may consider. For example, factors such as neighborhood quality, local reputation, future infrastructure, regulation, and market sentiment are important but harder to measure. Another limitation is that the visual illustration focuses on selected New York City home values, while the model itself is tested at the state level.

Although New York City is useful for showing how one major real estate market changes over time, it does not capture the full variation within the U.S. housing market. Future research could extend the analysis to more granular city-, neighbourhood-, or property-level data to test whether the model performs differently across local real estate markets. Machine learning models also depend on past data. This means they may work well when the future is similar to the past, but they may fail when market conditions suddenly change.

Further research should test the models during different periods, such as before, during, and after crises, to understand when AI performs well and when it fails. Another limitation is that comparing AI directly with human investors is difficult. Human investors use judgement, experience, and local knowledge that may not appear in the dataset. Further research could compare AI predictions with real investment decisions made by professional real estate investors or funds.

These limitations show that future research should use larger datasets, include more risk factors, and compare AI predictions with actual human investment decisions. This would help better answer when AI can improve real estate investing and when human judgement is still needed.

CONCLUSION

The findings of this research suggest that machine learning approaches offer an improvement in the house-price forecast over the simple benchmark but do not beat traditional statistical modeling consistently. Random Forest made more precise forecasts than the no-change benchmark in all states, proving that this algorithm was able to identify important factors other than the assumption of the stability of house prices. Nevertheless, OLS with state fixed effects demonstrated better performance in 48 of 51 states and had the lowest MAE and MAPE, while Random Forest showed the lowest RMSE due to the reduction of some extremely inaccurate forecasts. The best model according to each evaluation criterion differs. Also, the importance of features helps us understand how the Random Forest model makes its forecasts. More than 82% of the total importance of the features in this model comes from one-month and three-month ZHVI returns, which means that the momentum in house prices during the last few months

became the main factor. Thus, the movement in house prices in the previous period turned out to be more predictive than macroeconomic or housing market variables.

The findings of this research suggest that machine learning approaches offer an improvement in the house-price forecast over the simple benchmark but do not beat traditional statistical modeling consistently. Random Forest made more precise forecasts than the no-change benchmark in all states, proving that this algorithm was able to identify important factors other than the assumption of the stability of house prices. Nevertheless, OLS with state fixed effects demonstrated better performance in 48 of 51 states and had the lowest MAE and MAPE, while Random Forest showed the lowest RMSE due to the reduction of some extremely inaccurate forecasts. The best model according to each evaluation criterion differs. Also, the importance of features helps us understand how the Random Forest model makes its forecasts. More than 82% of the total importance of the features in this model comes from one-month and three-month ZHVI returns, which means that the momentum in house prices during the last few months became the main factor. Thus, the movement in house prices in the previous period turned out to be more predictive than macroeconomic or housing market variables.

REFERENCES

Seagraves, P. (2024). Real Estate Insights: is the AI revolution a real estate boon or bane?. *Journal of Property Investment & Finance*, *42*(2), 190-199.

Mooya, M. M. (2016). Real estate valuation theory. *Springer Berlin Heidelberg. doi*, *10*, 978-3.

Mukhamediev, R. I., Popova, Y., Kuchin, Y., Zaitseva, E., Kalimoldayev, A., Symagulov, A., ... & Yelis, M. (2022). Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges. *Mathematics*, *10*(15), 2552.

Naranjo, A., & Ling, D. C. (1997). Economic risk factors and commercial real estate returns. *The Journal of Real Estate Finance and Economics*, *14*(3), 283-307.

Lin, Z., & Vandell, K. D. (2007). Illiquidity and pricing biases in the real estate market. *Real Estate Economics*, *35*(3), 291-330.

Kohlscheen, E. (2025). Forecasting house prices. *arXiv preprint arXiv:2509.21460*.

Wang, Y. (2021, February). House-price prediction based on OLS linear regression and random forest. In *2021 2nd Asia Service Sciences and Software Engineering Conference* (pp. 89-93).