Intelligent Real Estate Advisor - Final Report - Team 114

P. Venieris, A. Colombi, J.A. Saint Antonin, R. Graeser, H. Tiruneh, M. Tekman

INTRODUCTION

The real estate market is subject to continuous fluctuations influenced by economic, demographic, and location factors. Accurate forecasting is essential for stakeholders to make informed decisions and maximize value of their investments. This project applies machine learning methodologies to develop predictive models and presents findings through an interactive visualization, ensuring accessibility and clarity in market trend analysis to end users.

PROBLEM DEFINITION

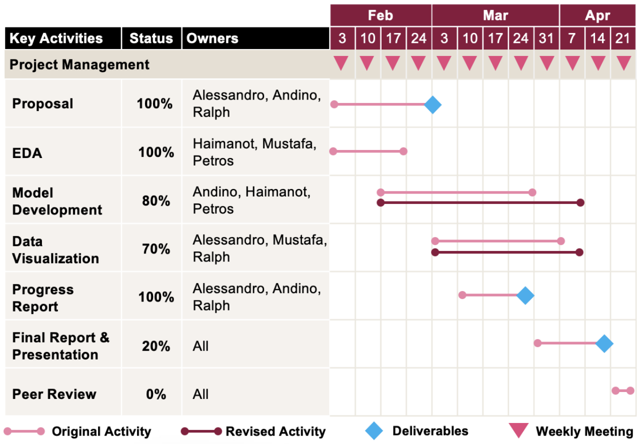
Access to reliable real estate information and price forecasts for investors can be challenging. To enable stakeholders to make more informed decisions regarding their real estate investments, this project addresses two primary objectives:

- Identifying areas for investment based on user-defined criteria through classical boolean filters, sorting, variable selection and visualization in a dashboard
- Predicting which regions are more likely to experience price increases.

To achieve these objectives, we integrate an intuitive interactive dashboard with a real estate prediction model. The predictions are based on machine learning algorithms and synthesize data from multiple sources to forecast anticipated price changes across counties in the United States. The resulting insights are presented in an accessible and user-friendly format through the interactive dashboard.

LITERATURE SURVEY

Real estate price forecasting involves predicting future property values primarily based on historical trends, typically at a regional or market level. This differs from real estate appraisal, which estimates the current value of an individual property based on its features, such as the number of rooms, size, amenities, and so forth Geerts (2023) [9] and Krause and Lipscomb (2016) [15]. While appraisal models use hedonic regression approaches, forecasting models rely more on time-series data. Advancements in machine learning, time-series forecasting, and spatial-economic integration have significantly increased the accuracy, demand, and utility of such dynamic forecasts and solutions (Gude, 2024 [10]; Sharma et al., 2024 [20]; Liu, 2023 [16]).



Key Activities	Status	Owners
Project Management		
Proposal	100%	Alessandro, Andino, Ralph
EDA	100%	Haimanot, Mustafa, Petros
Model Development	80%	Andino, Haimanot, Petros
Data Visualization	70%	Alessandro, Mustafa, Ralph
Progress Report	100%	Alessandro, Andino, Ralph
Final Report & Presentation	20%	All
Peer Review	0%	All

Figure 1: Gantt diagram of the work plan. All team members have contributed a similar amount of effort.

Traditional house price prediction often relies on regression-based approaches like linear, ridge, and LASSO regression. While effective for appraising current property values, these models struggle to forecast future trends [12], [7] [17] [6], especially in dynamic markets. Liu (2023) [16] found LASSO superior to linear regression, achieving values R^2 of 0.9 but requiring careful hyperparameter tuning to avoid overfitting. Similarly, Wijono et al. (2024) [25] reported that random forest outperformed linear regression (R^2 = 0.73 vs. 0.41), further highlighting the limitations of simpler models (Wijono et al., 2024) [25]. Dhar and Manikandan (2023) [21] emphasized LASSO's value for feature selection but noted its inability to adapt to economic fluctuations, socioeconomic shifts, supply and demand (Dhar & Manikandan, 2023) [21].

Time-series models like ARIMA have proven effective for short-term predictions but struggle with long-term volatility. Wijono et al. (2024) [25] Liu (2023) [16] and Belej (2015) [1] demonstrated the effectiveness of ARIMA and support vector regression for short-term forecasts, achieving high accuracy within limited temporal windows. Hurley and Sweeney (2022) [14], and Tchuente and Nyawa (2022) [23], applied spatial clustering in Ireland and France respectively, showing that forecasts within homogenous regions were more reliable. Özöğür Akyüz et al. (2023) [26] highlighted the effectiveness of hybrid models, combining time-series approaches with spatial clustering for regional forecasts.

Ensemble models like XGBoost, LightGBM, and random forest have emerged as powerful tools for real estate price forecasting [19]. These models combine multiple weak learners to improve prediction accuracy and handle complex, non-linear relationships. Guliker et al. (2022) [11] found XGBoost explained 83% of Dutch price variance, outperforming traditional hedonic models. Liu (2023) [16] and Wijono et al. (2024) [25] reported up to 94% R^2 using ensemble methods with temporal and spatial features. Rico-Juan and Taltavull de La Paz (2021) [18] highlighted that ensemble models also provide better transparency and interpretability when combined with spatial analytics.

Although many studies focus solely on historical prices, recent research highlights the importance of macroeconomic indicators and spatial factors for medium to long-term forecasts. Kameni et al. (2023) [24], Folmer et al. (2022) [11], and Wijono et al. (2024) [25] emphasized that variables like interest rates, unemployment, and regional economic health significantly influence housing market trends. Tekouabou et al. (2023) [24] and Chen et al. (2021) [8] argued that geospatial analytics further enhance model reliability by accounting for regional variations.

While advancements in real estate forecasting have improved accuracy through time-series models, ensemble learning, and spatial-economic analysis, these approaches predominantly remain fragmented. This project takes a more integrated approach, combining time-series forecasting with real estate indicators, such as days on market, alongside advanced machine learning models and spatial-economic insights. By synthesizing these elements, we aim to provide more dynamic, US county-level forecasts, better reflecting the complex interactions that drive housing markets. This approach has the potential to enhance decision-making for investors, policymakers, and industry stakeholders in an increasingly volatile market landscape.

METHODS

As discussed in the problem description, the project consists of two major steps:

- Dashboard design: where human/computer interaction know-how is used to implement an intuitive, targeted dashboard to assist users.
- Data and Computation: sourcing and cleaning of the data, feature engineering and model cross validation and testing.

We will begin with a discussion about why we believe our approach may be better than the state of the art and then dedicated sections to cover the two items above. Experiments and test results are in a later section.

Our approach vs state-of-the-art

The project innovations compared to the existing literature (to the best of our knowledge) are:

- Combining open real estate data with socio-economic indicators to enhance contextual understanding of housing markets
- Advanced feature engineering to maximize prediction accuracy.
- Highly granular short-term modeling (1,900 county-specific machine learning models) for localized accuracy coupled with robust mid term modelling (up to one year predictions)
- making high quality forecasts openly available to non-institutional investors.
- Interactive and intuitive dynamic dashboard with real-time predictions and comparative analysis tools.

Best ideas of the project are, in our opinion:

- Supporting real estate investment decisions with state-of-the art machine learning
- Highly localized models that provide better accuracy
- Democratizing access to high quality information through by publishing openly in the internet.

Although this is just an initial effort (no capital, short and part time effort, see Figure 1), we believe this

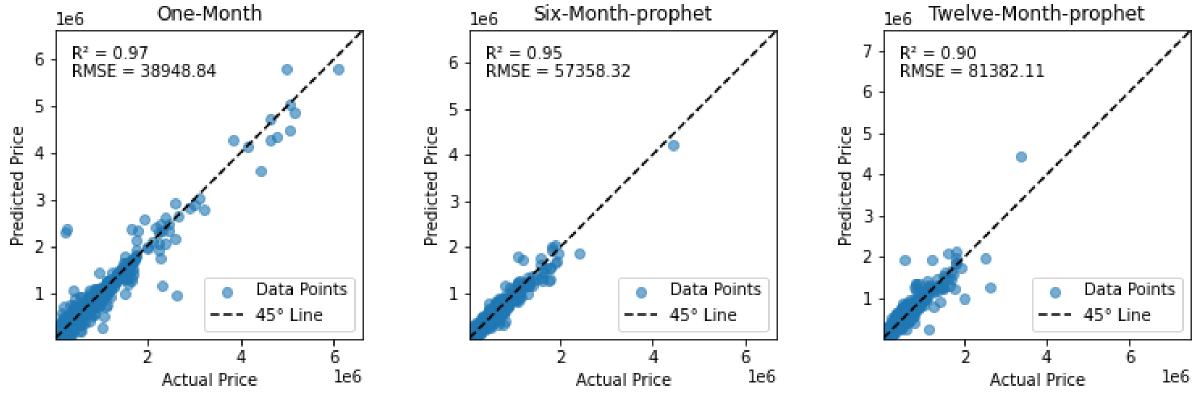


Figure 2: Actual vs Predicted price in test set for best 1-month and 6-month and 12-month Prophet models

type of tool can contribute significantly to leveling the playing between corporate investors and private individuals.

Computation Methods

This subsection introduces a description of the methods utilized to create forecasts of the price changes in housing prices. There are two sources of data:

- RedFin data which is a comprehensive collection of real estate data, including information on homes for sale, sold properties, market trends, and related analytics. It provides data on housing market trends, including information on supply and demand [5].
- American community Survey (run by the US Census Bureau) contains socio-economic indicators such as population size and unemployment, poverty and graduation rates [3].

Data Preparation: Data was constrained to the last 5 years as initial testing suggested previous years might even have a negative impact on the models because they no longer represent our current market conditions (due to COVID).

The initial intention was to use the redfin data in its full scope (up to zip code data) which amounted to 3GB of data. However, we found that zip code level data was sparse, making time series predictions very difficult. Further, we could not find socioeconomic data to match the zip code data density. Therefore we resorted to make predictions at a county level, which had continuous redfin data and socio-econonmic information was available.

Data cleaning consisted in removing counties with less than 48 months of data in this 60 month period. The very few missing data was imputed by interpolation using dates as an independent variable. Also, over 1000 homes sold over the 60 months was required, as small sample sets could be dominated by unrepresentative data. Counties not meeting this criteria are excluded (no forecast will be available in the dashboard).

Multiple approaches to feature engineering were attempted (ratios, differences, etc) in order to enrich the original dataset. Among the new features, the most useful appear to be related to inventory (homes_sold over inventory) an delays in the selling of houses (new_listings inventory). Socio economic data had yearly frequency but it was interpolated to match the monthly information density of the RedFin data.

Predictions of price were very successful, achieving results of R^2 ranging between 0.97 and 0.9 depending on the time horizon of the prediction, which is on a par if not better than the literature surveyed (see figure 2) However, we aimed to predict the more ambitious year on year price change, which is much more challenging. Figure 3 shows the prediction error in the blind test set for the best models for prediction of YOY price change using different time windows, ranging from 1 to 12 months. The results show RMSE of 13% to up to 22%, which is significant.

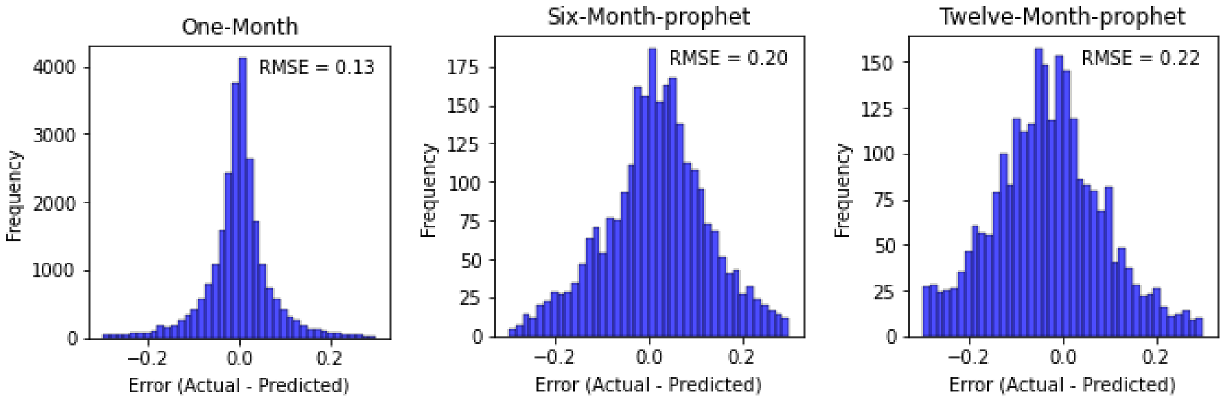


Figure 3: Error histograms for best models in each time horizon

In Figure 4 we see actual vs prediction YOY price change. These results suggest this type of model should only be used on portfolio decisions as individual predictions may contain high error. A detailed discussion of the models used is presented below.

Machine Learning Models. Based on the literature review, the following models were trained:

- Ridge regression with L2 regularization with engineered features, such as lagged sale prices, moving averages, and percentage changes. This regularization reduces overfitting and multicollinearity [13].
- XGBoost, which employs a temporal train-test split alongside engineered features. Weak learners (shallow trees) iteratively improve predictions by correcting errors from previous iterations. Built-in regularization ensures robustness, enabling the model to effectively capture complex time-series patterns [4].
- Random Forest Regression which trains an ensemble of decision trees on a time-based split. Each tree predicts the median sale price independently, and the final output is the average of these predictions (YoY price change is calculated outside). This method captures non-linear relationships and interactions in the data. This type of model captures complex time-dependent patterns [2].
- Prophet: an open-source forecasting tool developed by Facebook. It is specifically designed for time series data and excels at handling seasonality, holidays, and missing data. Prophet uses an additive model to decompose time series into trend, seasonality, and holiday effects, making it highly interpretable. It is robust to outliers and works well with irregularly spaced data [22]

For Random Forest and XGBoost, 4 fold cross-validation is performed to determine the best model hyper-parameters prior to testing. Similarly, cross validation was used in prophet to determine optimal model parameters.

All the models were used forecast with time windows of one, three, six and twelve months. The best models were selected in each case. For the one and three month periods, the best model was selected for each county and the results were quite reliable, expecially with XGBoost. As for the longer time lags of six and twelve months, we found that prophet was outperforming the other options.

Visualization Methods

The dashboard provides an interactive, data-driven representation of real estate market dynamics at the county level. It includes a customazible tool tip, search, sort, and filter functions for comparing counties based on property type, price range, YoY price change and other economic indicators. The primary visualization is a Choropleth map (see Figure 5) that displays expected YoY price change upon

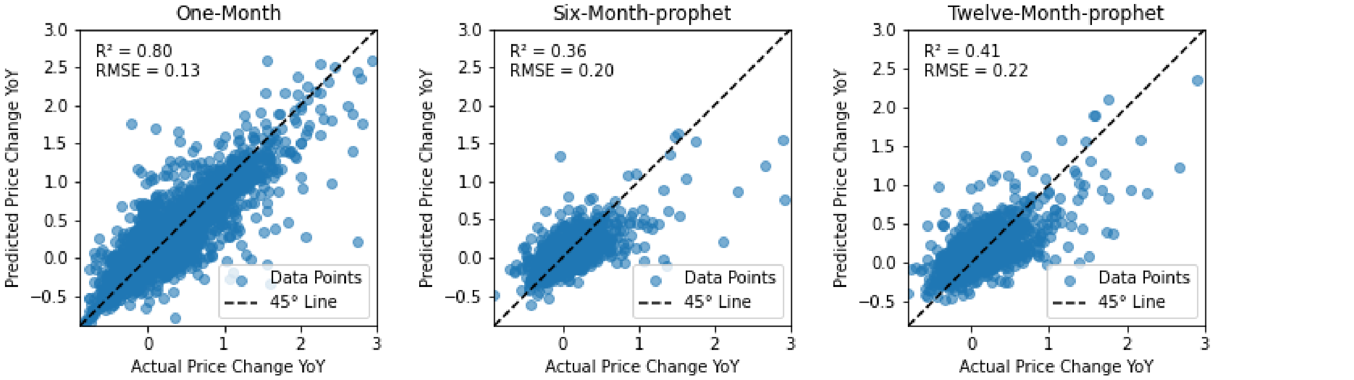


Figure 4: Actual vs Predicted - YoY price change on test set for 1, 6 and 12-month lag models

hover. Other variables (e.g. sales to list ratio, inventory levels, months of supply) can be displayed in table format.

A pass-through filtering feature in the analysis dashboard allows users to explore and compare specific counties in detail through time-series plots of selected indicators like average sale prices, unemployment levels, population growth or inventory levels.

The dashboard was developed in Power BI for its advanced visualization capabilities and interactivity. A custom U.S. county shapefile (from the U.S. Census Bureau, converted for Power BI) drives the geographic view. Geographic name normalization was performed to ensure accurate map joins. All charts are dynamically linked, with tooltips and table variables synchronized via parameterized filtering. Multi-variable selection and a responsive layout (using a PowerPoint-designed template) enhance usability and visual presentation.

Predictive price models are deployed via csv tables, as their update can at most be done monthly, so real-time capabilities are not required. Updates can be programmed monthly as new data becomes available.

Overall, the visualization's strength lies in its integration of predictive analytics, customizable KPIs, and synchronized interactivity. It provides a unified tool for both opportunity identification and due diligence – a relatively rare feature in public real estate dashboards.

EVALUATION AND EXPERIMENTS

Testing the models

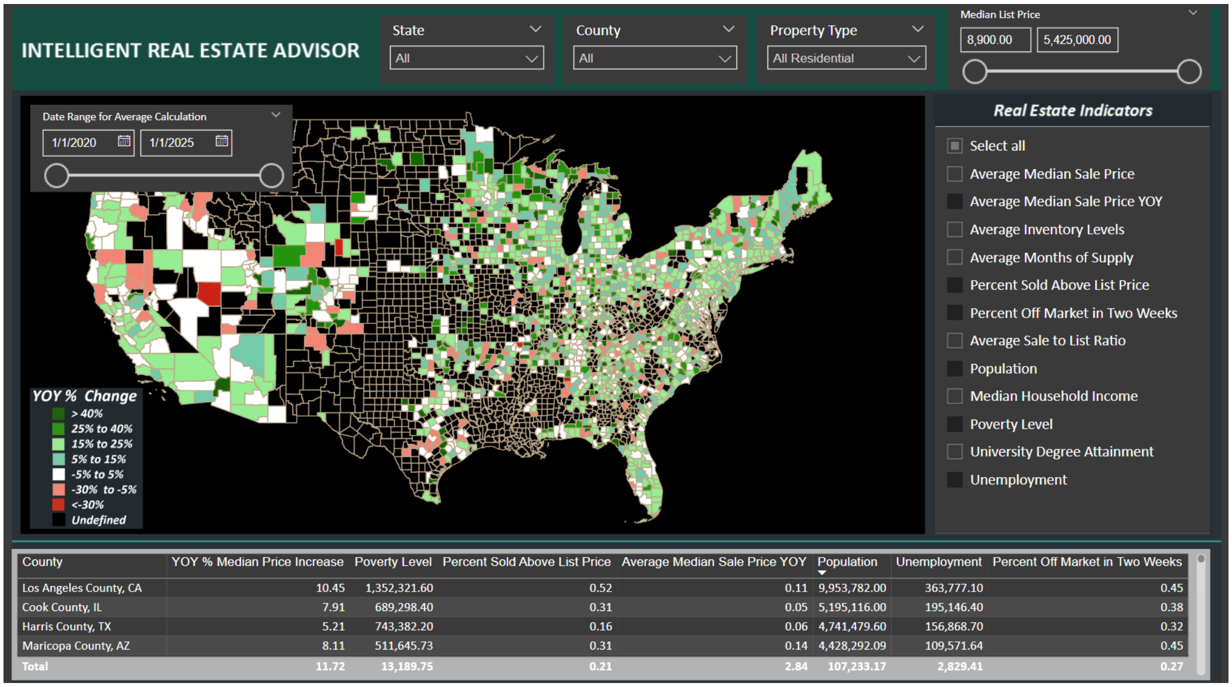
The methods section described how models were tested and the results. Data was split into cross-validation and testing sets using a time-aware split, where historical observations form the cross-validation set and recent data is reserved for testing. In our case in particular, the last year was used for testing and the four prior years are used for cross validation.

Features (e.g., lagged values, moving averages, and percentage changes) were derived solely from training data and applied consistently to the test set. This setup prevented data leakage and ensured realistic evaluation (further investigation into feature engineering is ongoing). Performance metrics, namely RMSE, and R^2 were used to assess the true forecasting capabilities, therefore ensuring it generalizes to unseen, temporally ordered data. Testing results were presented in Figure 4 and 3 and discussed in the methods section.

Although our predictions on future prices were highly accurate (on a par if not better than most efforts in the literature), prediction price change YoY turned out to be much more challenging. The testing methodology faithfully reflects the use case of the predictions

Dashboard experimentation

The Dashboard was tested in use by team members that were not part of the design and implementation. Multiple feedback sessions were conducted to refine



State: All

County: All

Property Type: All Residential

Median List Price: 8,900.00 – 5,425,000.00

Date Range for Average Calculation: 1/1/2020 – 1/1/2025

■ Select all
□ Average Median Sale Price
■ Average Median Sale Price YOY
□ Average Inventory Levels
□ Average Months of Supply
■ Percent Sold Above List Price
■ Percent Off Market in Two Weeks
□ Average Sale to List Ratio
■ Population
□ Median Household Income
■ Poverty Level
□ University Degree Attainment
■ Unemployment

County	YOY % Median Price Increase	Poverty Level	Percent Sold Above List Price	Average Median Sale Price YOY	Population	Unemployment	Percent Off Market in Two Weeks
Los Angeles County, CA	10.45	1,352,321.60	0.52	0.11	9,953,782.00	363,777.10	0.45
Cook County, IL	7.91	689,298.40	0.31	0.05	5,195,116.00	195,146.40	0.38
Harris County, TX	5.21	743,382.20	0.16	0.06	4,741,479.60	156,868.70	0.32
Maricopa County, AZ	8.11	511,645.73	0.31	0.14	4,428,292.09	109,571.64	0.45
Total	11.72	13,189.75	0.21	2.84	107,233.17	2,829.41	0.27

Figure 5: Interactive Dashboard for real estate price prediction

the functionality and interactivity of the dashboard. Adjustment were made to the layout design, color scales, displayed variables and many other aspects. The team experimented with multiple variables on the map but chose to display only forecasted YoY price change for simplicity and due to its importance for decision support.

Many experiments were carried out with different base map types within Power BI to best represent the county level layer.

The final dashboard can be observed in Figure 5. The peer review sessions served to maximize the useful information available to users when assessing real estate investments.

CONCLUSIONS AND DISCUSSION

With minimal resources and open source data we have implemented a tool that supports real estate investment decisions with price change forecasts, supply/demand, economic data and ample functionality to search for opportunities suitable for the users. It is all packaged in an intuitive and easy-to-use dashboard that combines forecasting, screening, and detailed analysis in a seamless user experience.

Future extensions of the present work could focus on the following:

- enriching the feature space to improve predictive model reliability
- extending granularity to zip level in the USA
- extend to regions where reliable data is available, such as Europe, Canada, Australia, etc.
- collect feedback from users to improve the dashboard functionality

Our hope is that, although forecasts on price changes are not highly accurate, they nevertheless can contribute to significant improvement on portfolio performance of non-institutional investors, due to the open source nature of the work.

Contribution statement

All team members have contributed a similar amount of effort (particularly during the literature research). Then tasks were split to accelerate progress: P. Venieris and H. Tiruneh focused on the price prediction, A. Colombi and J.A. Saint Antonin oversaw the reporting and coordination efforts, R. Graeser and M. Tekman jointly designed and implemented the dashboard.

REFERENCES

[1] Miroslaw Belej and Sławomir Kulesza. 2015. The Dynamics of Time Series of Real Estate Prices. *Real Estate Management and Valuation* 23, 4 (2015), 35–43. https://doi.org/10.1515/remav-2015-0034

[2] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[3] United States Census Bureau. 2025. American Community Survey (ACS) Data. https://www.census.gov/programs-surveys/acs/data.html. Accessed: 2025-03-17.

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794. https://doi.org/10.1145/2939672.2939785

[5] Redfin Corporation. 2025. Redfin Data Center. https://www.redfin.com/news/data-center/. Accessed: 2025-03-17.

[6] Jason Robert Bailey et al. 2022. Hedonic Models of Real Estate Prices: GAM Models; Environmental and Sex-Offender-Proximity Factors. *Journal of Risk and Financial Management* (2022).

[7] Jiawei Wang et al. 2023. Understanding Negative Equity Trends in U.S. Housing Markets: A Machine Learning Approach to Predictive Analysis. *Journal of Economics, Finance and Accounting Studies* 5, 6 (2023), 99–120. https://doi.org/10.32996/jefas.2023.5.6.10

[8] Ming-Hsiu Chen et al. 2021. Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism. *IEEE Access* 9 (2021), 55244–55259. https://doi.org/10.1109/ACCESS.2021.3071306

[9] M. Geerts, S. Broucke, and J. De Weerdt. 2023. A Survey of Methods and Input Data Types for House Price Prediction. *International Journal of Geo Information* 12, 200 (2023). https://doi.org/10.3390/ijgi12050200

[10] V. Gude. 2024. A multi-level modeling approach for predicting real-estate dynamics. *International Journal of Housing Markets and Analysis* 17, 1 (2024), 48–59. https://doi.org/10.1108/IJHMA-02-2023-0024

[11] E. Guliker, E. Folmer, and M. van Sinderen. 2022. Spatial Determinants of Real Estate Appraisals in The Netherlands: A Machine Learning Approach. *ISPRS International Journal of Geo-Information* 11, 2 (2022), 125–. https://doi.org/10.3390/ijgi11020125

[12] R. Gupta, H.A. Marfatia, and C. Pierdzioch et al. 2022. Machine Learning Predictions of Housing Market Synchronization across US States: The Role of Uncertainty. *Journal of Real Estate Finance and Economics* 64 (2022), 523–545. https://doi.org/10.1007/s11146-020-09813-1

[13] Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12, 1 (1970), 55–67. https://doi.org/10.1080/00401706.1970.10488634

[14] A.K. Hurley and J. Sweeney. 2024. Irish Property Price Estimation Using A Flexible Geo-spatial Smoothing Approach: What is the Impact of an Address? *J Real Estate Finan Econ* 68 (2024), 355–393. https://doi.org/10.1007/s11146-022-09888-y

[15] Andy Krause and Clifford A. Lipscomb. 2016. The Data Preparation Process in Real Estate: Guidance and Review. *Journal of Real Estate Practice and Education* 19, 1 (2016), 15–42. https://doi.org/10.1080/10835547.2016.12091756

[16] Z. Liu. 2023. Real Estate Price Prediction based on Supervised Machine Learning Scenarios. *Highlights in Science, Engineering and Technology* 39 (2023), 731–737. https://doi.org/10.54097/hset.v39i.6637

[17] Ping-Feng Pai and Wen-Chang Wang. 2020. Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices. *Applied Sciences* 10, 17 (2020), 5832. https://doi.org/10.3390/app10175832

[18] Juan Ramón Rico-Juan and Paloma Taltavull de La Paz. 2021. Machine Learning with Explainability or Spatial Hedonics Tools? An Analysis of the Asking Prices in the Housing Market in Alicante, Spain. *Expert Systems with Applications* 171 (2021), 114590. https://doi.org/10.1016/j.eswa.2021.114590

[19] H. Sharma, H. Harsora, and B. Ogunleye. 2024. An Optimal House Price Prediction Algorithm: XGBoost. *Analytics (Basel)* 3, 1 (2024), 30–45. https://doi.org/10.3390/analytics3010003

[20] S. Sharma, S. Kumari, S. Goyal, and R. Nirala. 2024. A Review: Real Estate Price Prediction using Machine Learning with Research and Trends. In *Proceedings of the 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*. 1239–1244. https://doi.org/10.1109/IC2PCT60090.2024.10486480

[21] Dhar T and Manikandan P. 2023. A Literature Review on Using Machine Learning Algorithm to Predict House Prices. *International Research Journal on Advanced Science* 5 (March 2023). Issue 05S. https://doi.org/10.47392/irjash.2023.S017

[22] Sean J. Taylor and Benjamin Letham. 2017. Forecasting at Scale. *PeerJ Preprints* 5 (2017), e3190v2. https://doi.org/10.7287/peerj.preprints.3190v2

[23] D. Tchuente and S. Nyawa. 2022. Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research* 308 (2022), 571–608. https://doi.org/10.1007/s10479-021-03932-5

[24] S. Tekouabou, S. Gherghina, E. D. Kameni, Y. Filali, and K. I. Gartoumi. 2023. AI-Based on Machine Learning Methods for Urban Real Estate Prediction: A Systematic Survey. *Archives of Computational Methods in Engineering* 31 (2023), 1079–1095. https://doi.org/10.1007/s11831-023-10010-5

[25] D. S. Wijono, F. M. Lienardi, J. N. Lisapaly, I. S. Edbert, and D. Suhartono. 2024. Regression Models with Different Levels of Complexity for Real Estate Price Prediction. In *Proceedings of the 2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*. 1–5. https://doi.org/10.1109/AIMS61812.2024.10512702

[26] S. Özöğür Akyüz, B. Eygi Erdogan, Ö. Yıldız, and P. Karadayı Ataş. 2023. A Novel Hybrid House Price Prediction Model. *Computational Economics* 62, 3 (2023), 1215–1232. https://doi.org/10.1007/s10614-022-10298-8