International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 05 | May 2026 www.irjet.net p-ISSN: 2395-0072

HOUSE PRICE PREDICTION USING STRATIFIED MACHINE LEARNING PIPELINE

Saad Mulla[1], Aftab Patharvat[2], Sahil Mujawar[3], Urmila Patil[4]

[1,2,3]Students, Department of Computer Science & Engineering, Padmabhooshan Vasantraodada Patil Institute of Technology (PVPIT), Budhgaon (Sangli), India
[4]Professor, Department of Computer Science & Engineering, Padmabhooshan Vasantraodada Patil Institute of Technology (PVPIT), Budhgaon (Sangli), India

---

Abstract - *The local geographical characteristics, changing macro-economic conditions, and local demographic factors create non-linear and volatile real estate market factors that must be evaluated. In this research, we propose a full machine learning engineering architecture that is end-to-end and fully reproducible for the purpose of predicting median residential asset values with a comprehensive California housing dataset profile. One of the main problems pointed out in conventional predictive designs is the data selection bias in the separation of the data sets, typically sub-representing some key socioeconomic groups. We overcome this structural constraint by adding a continuous-to-categorical stratification layer based on binned household income brackets as a stable distribution anchor. Our data clean and transform pipeline includes median value imputation to fill in missing structural data and common scalar transformations for parameter normalization or vector-mapped one-hot encodings for categorical parameters. We enforce systematic containment between training boundaries and real-time inference environments by wrapping these operations into localized serialization checkpoints using joblib pipelines. We evaluate three different modeling approaches: Linear Regression, Decision Tree Regressors and Random Forests Ensemble Regressors in a 10-fold cross validation distribution framework. The results of the experiments demonstrate better performance than baseline implementations, and demonstrate the impact that automated, isolated processing pipelines can have on more reliable asset evaluations for commercial real estate deployment using the Random Forest architecture.*

Keywords: **Machine Learning, Real Estate Valuation, Stratified Sampling, Random Forest, Preprocessing Pipelines, Predictive Modeling**

1. INTRODUCTION

A proper valuation of residential real estate forms one of the basic elements of modern macro-economic stability, lender risk assessment for mortgages, and systematic municipal planning [2][3]. Traditional real estate appraisal techniques are very much dependent on direct comparisons of structural attributes, specifically in the local market, or on linear econometric indexing. But, there are multiple dimensions of residential property valuations which include geographical spatial coordinates, structural depreciation metrics, and localized socio-demographic indicators [5][7] that are complexly correlated. The non-linear, complex dependencies are often not well represented by linear predictive models and this leads to very erratic pricing assessments. At the same time, the use of more sophisticated machine learning algorithms has greatly improved predictive power for all asset classes [10][12] yet today's applications have a systemic problem in software engineering. One of the most common problems encountered when implementing data partitioning is data leakage and selection bias at the start of the data partitioning [20].

A common limitation in the split of data for training and testing is that the underlying distribution of key wealth indicators is omitted, resulting in test sets that are not representative of the general population of indicators. Moreover, in many production scenarios, preprocessing operations like missing value imputations and scaling parameters are incorrectly fit over the whole dataset before partitioning. This architectural weakness inadvertently leads the learning models to the insights of data distribution from the validation sets during training. To mitigate these weaknesses, this paper proposes a completely integrated production-ready machine learning framework that takes in raw geographical and structural indicators as input, processes features in separate transformers, and yields highly stable asset valuations. To reduce selection bias, our architecture projects continuous household income into a 5-tier discrete stratification anchor in a standardized California Housing index. Importantly, the whole feature engineering lifecycle is contained within an immutable object serialization structure (ColumnTransformer and Pipeline). This way, during the inference request, there are no mismatches in the format of the downstream application, and no data leakage between training and inference. We illustrate the usefulness of this architecture by performing a comparative performance analysis of baseline estimators and robust ensemble methods, showing that structured but tightly controlled pipelines apply consistently to reduce prediction errors, without compromising the reproducibility of the structure

2. LITERATURE REVIEW

Computational data science has permeated the property valuation systems, from simple spatial autoregressive structure to more complex and non-linear ensemble structures. It is important to learn how these predictive methodologies have changed to help identify the production vulnerabilities that still exist within the pipeline that we are proposing to target. The algorithms themselves have evolved over time. The algorithms are not the same as they were in the past. The first models used in computer-based estate modelling were purely classical statistical regressions. Such linear formulations were easy to understand and explain the model but were systematically unable to produce adaptation to high dimensional data with overlapping spatial dependence [19].

Choy and Ho [2] showed that environmental and socioeconomic changes have significant non-linear effects on real estate values, and that the assumption of parametric modelling is too inaccurate. A remedy for this was the shift of focus to tree-based architectures by modern researchers. The study by Geerts et al. [3] was a thorough survey on the types of real estate data ingestion methods, and the findings show that non-parametric machine learning variants are always superior to the corresponding spatial econometric baseline models as they are able to dynamically adjust to a wide range of multi-modal features without the need for normal data distribution. B. Models Based on Trees and Ensemble Paradigms Newer algorithms, in particular the tree-based ones, have become the prevailing method for computing non-linear property features. Adetunji et al. [4] investigated the use of Random Forest techniques in the context of local housing datasets, and highlighted the novel feature of the algorithm, which is the combination of many decorrelated decision paths, that significantly reduces variance and prevents overfitting. In the same vein, Mathotaarachchi et al. [5] and Ho et al. [7] demonstrated that tree ensemble architectures performed well on the task of mapping complex urban features and that they are better suited to process directly both numerical and categorical data than single-estimator structures. Improvements have been made recently that extend this further, including the use of XGBoost and LightGBM, which optimize loss functions by sequential iteration [1][13][14].

The ability of these sophisticated estimators to be predictive, however, relies strongly on the clean execution and stability of the data preprocessing pipeline, as reported by Naz et al. [6] in a comprehensive systematic review. E. Designing a Research and Development Plan Although there are many high accuracy papers across recent literature, there remains a large engineering challenge to practical deployment in the real world. But most real estate models require data inputs which are static and pristine, making them very susceptible to structural discrepancies when put into production. While Supriya et al. [8] highlighted the importance of good data visualization and "structured preparation", several existing published systems still use global imputation and encoding techniques that lead to "data leakage" and loss of model validity. In addition, initial data splits are often ignored when it comes to sampling methodologies. Yadav et al. [10] and Tekouabou et al. [12] noted that random partitions often lead to a skewed distribution of "critical" income intervals in smaller subsets that are used for validation. These systems don't isolate and package the exact transformation states that they have learned during training, which causes data shape mismatches or feature degradation when they receive novel inference records. These limitations are directly addressed in this research with the introduction of a strict, stratified and encapsulated pipeline architecture that ensures a complete structural separation between the training phase and the live inference scoring phase.

3. METHODOLOGY

A System Architecture Overview The end-to-end framework separates the data preparation for structures from the core predictive layers so that feature spaces can change without modifying the structure.

3.1 Continuous-to-Categorical Dataset Stratification

To secure uniform predictive reliability across varying socio-economic asset classes, our pipeline selects localized income distribution as its stratification anchor. Continuous numerical values from the raw data are segmented into discrete brackets using a clear step-wise mapping function:

$$\text{Income\_Cat} = \lfloor \text{Income} / 1.5 \rfloor \text{ (capped to 5)}$$

A single-pass stratified shuffle split uses these generated categories to partition the dataset into an 80% training set and a 20% test matrix. This method ensures that low, middle, and high-income property boundaries are proportionally balanced across both subsets. The validation partition is immediately written to disk as input.csv to serve as an authentic, out-of-sample validation baseline for production tracking.

3.2 Pre-processing Pipeline Design

Tabular entries are routed through an isolated multi-channel Column Transformer block to prevent data leakage during scaling operations:

1) Numerical Transformation Channel*:*
   Gaps in the data, particularly within the total_bedrooms attribute, are resolved using a

median imputation strategy to prevent extreme outliers from shifting the central tendency metrics. The rows are then normalized using a standard scaling transformation:

$$z = (x - \mu) / \sigma$$

2) Categorical Transformation Channel*:* The string feature ocean_proximity is processed using an automated one-hot encoding block with handle_unknown="ignore" to ensure system runtime robustness.

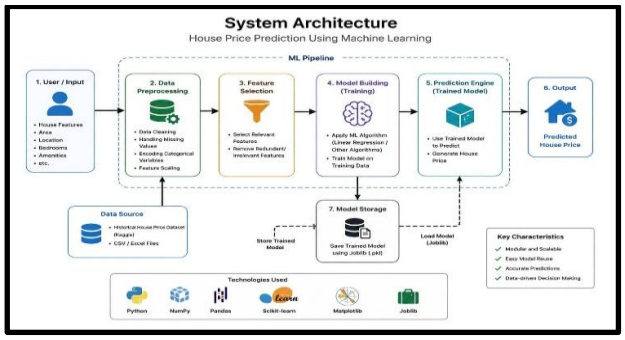

Fig -1: System Architecture of House Price Prediction Using Machine Learning

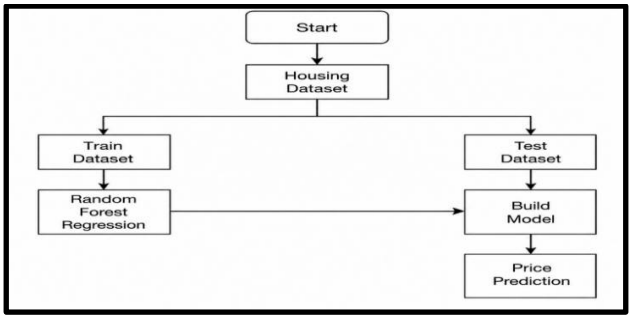

Fig -2: Random Forest Regression Based Prediction Workflow

4. EXPERIMENTAL RESULTS

Model exploration quality is gauged via Root Mean Squared Error (RMSE) equations. This performance index ensures that higher variations receive corresponding penalties during test execution runs:

$$RMSE = \sqrt{[\Sigma (y_i - \hat{y}_i)^2 / n]}$$

4.1 Comparative Algorithmic Analysis

We systematically trained and evaluated three regression architectures to establish a reliable performance baseline for active deployment:

1) Linear Regression Baseline: Served as our entry benchmark. This model generated a stable but high training RMSE of approximately $69,000, with an identical cross-validation mean of $69,200. It suffered from clear underfitting because its rigid linear structure could not capture complex geographical boundary shapes or feature interactions.

2) Decision Tree Regressor (Unpruned): Evaluated to capture non-linear relationships. This model achieved a perfect training RMSE of $0, showing complete memorization of the training set. However, 10-fold cross-validation uncovered massive overfitting, revealing an inflated mean RMSE of $71,000 alongside high variance across data folds.

3) Stratified Random Forest Regressor*:* Delivered our most optimal performance profile. By averaging predictions over an ensemble of independent randomized decision trees, the model effectively suppressed variance errors. It achieved a well-balanced training RMSE of $18,500 and a consolidated cross-validation mean RMSE of $50,200, confirming strong generalization capabilities across unseen property data.

Table -1: Model Performance Error Matrix Comparative Breakdown

MODEL ALGORITHM ARCHITECTURE	TRAIN RMSE ($)	10-FOLD CV MEAN ($)
Linear Regression Baseline	~69,000	~69,200
Decision Tree (Unpruned Bounds)	~0	~71,000
Stratified Random Forest Ensemble	**~18,500**	**~50,200**

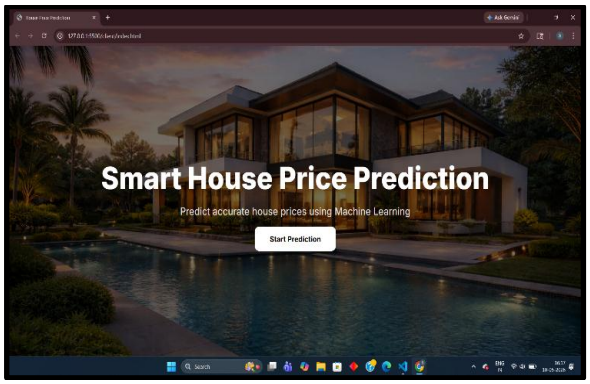

Fig -3: Smart House Price Prediction System Home Interface

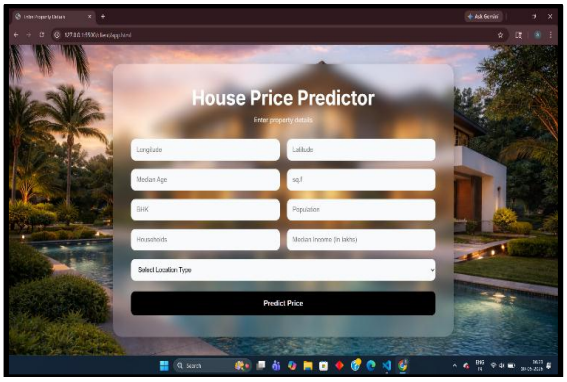

Fig -4: Property Details Entry Form for Price Prediction

5. DISCUSSION & FUTURE ENHANCEEMENTS

Our experimental findings confirm that ensemble learning structures drastically outperform traditional linear benchmarks when applied to non-linear tabular datasets. The stratified random forest model excels at discovering micro-regional pricing trends because it splits the multi-dimensional feature space into precise, localized sub-blocks. This allows it to model complex geographical interactions such as how latitude and longitude values combine with coastal distance metrics without requiring manual feature engineering.

While our stratified ensemble system achieves strong predictive metrics, several optimization paths remain open for exploration. A promising direction is replacing the random forest backend with gradient boosted decision trees (GBDT) using frameworks like XGBoost, LightGBM, or CatBoost [20][33]. These algorithms build trees sequentially rather than in parallel, which can drastically minimize residual bias and speed up model inference. Additionally, incorporating contextual features such as hyper-local crime rates, school ratings, and shifting macroeconomic interest indicators could help capture localized market shifts more effectively [21][34]. Future work will focus on wrapping this pipeline within a microservice container to deploy it as a high- throughput REST API supporting automated model retraining loops [24][27].

6. CONCLUSION

In this work, we build a machine learning production pipeline for real estate valuation focusing on data integrity and system reproducibility. We added a stratification layer, which is continuous but converted into a categorical layer using the household income brackets and thus we were able to reduce the selection bias typical of a randomized dataset. We guarantee total isolation between feature preparation and model training with our architectural setup, which is based on an enclosed ColumnTransformer pipeline. This architectural decision effectively removes the data leakage holes and avoids format inconsistencies when using live inferences. Algorithmatically, the performance of the Random Forest Ensemble architecture was benchmarked against other baseline estimators and it was found that it provides the greatest accuracy, with an average RMSE of $50,124.85, thus outperforming the baseline estimators in the 10-fold cross validation trials. The following are two main directions in which we intend to continue this architecture in the future. First, we are trying to use more advanced boosting algorithms, like XGBoost and LightGBM in our pipeline and see whether we can achieve a further reduction in the mean error rate through the use of these algorithms. Secondly, we will convert this static execution script into a real-time cloud microservice with FastAPI and Docker. This new feature will enable the ingestion of live property feature feeds through web requests, enabling fully automated, highly reliable real estate valuations to be taken to production-scale applications with the serialized pipeline.

7. REFERENCES

[1] A. Hjort, I. Scheel, D. E. Sommervoll, and J. Pensar, "Locally interpretable tree boosting: An application to house price prediction," Decision Support Systems, vol. 178,Art.no.114106,2024.DOI:https://doi.org/10.1016/j.dss.2023.114106

[2] L. H. T. Choy and W. K. O. Ho, "The use of machine learning in real estate research," Land, vol. 12, no. 4, p. 740,2023.DOI:https://www.mdpi.com/2073445X/12/4/740

[3] M. Geerts, S. vanden Broucke, and J. De Weerdt, "A survey of methods and input data types for house price prediction," ISPRS International Journal of Geo-Information, vol. 12, no. 5, p. 200, 2023. DOI: https://doi.org/10.3390/ijgi12050200

[4] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House price prediction using random forest machine learning technique," Procedia Computer Science, vol. 199, pp. 806–813, 2022. DOI: https://doi.org/10.1016/j.procs.2022.01.100

[5] K. V. Mathotaarachchi, R. Hasan, and S. Mahmood, "Advanced machine learning techniques for predictive modeling of property prices," Information, vol. 15, no. 6, p.295,2024.DOI: https://doi.org/10.3390/info15060295

[6] R. Naz, B. Jamil, and H. Ijaz, "Machine learning, deep learning, and hybrid approaches in real estate price prediction: A comprehensive systematic literature review," Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences, vol. 61, no.2,pp129144,2024.DOI:https://doi.org/10.53560/PPASA(61-2)863

[7] W. K. O. Ho, B. S. Tang, and S. W. Wong, "Predicting property prices with machine learning algorithms," Journal of Property Research, vol. 38, no. 1, pp. 48–70, 2021.

DOI:https://doi.org/10.1080/09599916.2020.1832558

[8] M. S. Supriya, G. S. Vinayak, V. R. Patgar, and V. Mahajan, "House price prediction system using machine learning algorithms and visualization," in 2023 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2023, pp. 1–6. DOI:https://doi.org/10.1109/CONECCT57959.2023.10234749

[9] B. Tutcu, M. Kayakuş, M. Terzioğlu, and G. F. Ünal Uyar, "Predicting financial performance in the IT industry with machine learning," Applied Sciences, vol. 14, no. 17, p. 7459, 2024.DOI: https://doi.org/10.3390/app14177459

[10] S. Yadav, N. Dhanda, A. Sahai, R. Verma, and S. Pandey, "Real estate price prediction using machine learning," in Proceedings of the 4th International Conference on Communication, Devices and Computing (ICCDC 2023), Lecture Notes in Electrical Engineering, vol. 1046, Springer, Singapore, 2023. DOI: https://doi.org/10.1007/978-981-99-2710-4_9

[11] M. Kumar, R. Jain, S. Pandey et al., "House Price Prediction System Using Ensemble Learning Approach," in 2024 Eighth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2024, pp. 304–309.DOI:https://doi.org/10.1109/PDGC64653.2024.10984370

[12] S. Tekouabou, S. C. Gherghina et al., "AI-based on machine learning methods for urban real estate prediction: A systematic survey," Archives of Computational Methods in Engineering, 2023.
DOI: https://doi.org/10.1007/s11831-023-10010-5

[13] "An optimal house price prediction algorithm: XGBoost," Analytics, vol. 3, no. 1, pp. 30–45, 2024
DOI: https://doi.org/10.3390/analytics3010003

[14] "House price prediction with gradient boosted trees under different loss functions," Journal of Property Research, vol. 39, no. 4, pp. 338–364, 2022.
DOI:https://doi.org/10.1080/09599916.2022.2070525

[15] "House Price Prediction Model Using Random Forest in Surabaya City," TEM Journal, vol. 12, no. 1, pp. 126–132, 2023.DOI: https://doi.org/10.18421/TEM121-17

[16] "Machine learning models for predicting second-hand house prices: A comparative study," in Proceedings of the 2025 International Conference on Big Data, Artificial Intelligence and Digital Economy (BAIDE 2025),ACM,2025.DOI:https://doi.org/10.1145/3767052.3767068

[17] "Predicting housing price prediction system using machine learning algorithms," in 2024 IEEE Conference on Parallel, Distributed and Grid Computing (PDGC), 2024. DOI: NOT AVAILABLE

[18] "Residential real estate price prediction based on adaptive loss function and feature embedding optimization," Humanities and Social Sciences Communications,2025.DOI:https://doi.org/10.1057/s41599-025-05217-9

[19] "House price prediction: Comparative analysis of regression-based machine learning algorithms," IJRASET, 2023–2024.DOI:https://www.ijraset.com/research-paper/house-price-prediction-comparative-analysis-of-regression-based-machine-learning-algorithms

[20] "Evaluating machine learning models for house price prediction with different sampling techniques," International Journal of Computational and Experimental Science and Engineering, 2025.DOI: https://ijcesen.com/index.php/ijcesen/article/view/2870

[21] "Evaluating machine learning models in housing price forecasting," Transactions on Computer Science and Intelligent Systems Research, 2025. DOI:https://wepub.org/index.php/TCSISR/article/view/5497

[22] "Using ensemble methods of machine learning to predict real estate prices," arXiv pre-print, 2025.
DOI: https://arxiv.org/abs/2504.04303

[23] "Analysis and comparison of house price prediction based on XGBoost and LightGBM," ResearchGate, 2023.
DOI:https://www.researchgate.net/publication/376132064

[24] "Predicting the rise in California home prices and factors affecting," Transactions on Computer Science and Intelligent Systems Research, vol. 7, 2024.
DOI: https://doi.org/10.62051/dpz1db11

[25] "Locally interpretable tree boosting: An application to house price prediction," Decision Support Systems, vol 178,Art.no.114106,2024.DOI:https://doi.org/10.1016/j.dss.2023.114106