International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2559 - 2563
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0381

Precision in Real Estate: Hybrid Multi-Stage Predictions for House Prices

Amudala Manasa[1], Pasupula Chandana[2], MD Shoaib Ahmed[3], Kumar Baradur[4]
[1,2,3]UG Scholar, Dept. of CSE-AIML, Sphoorthy Engineering College, Hyderabad, Telangana, India.
[4]Assistant professor, Dept. of CSE-AIML, Sphoorthy Engineering College, Hyderabad, Telangana, India.
Emails: *amanasa1406@gmail.com[1], chandanapasupula42@gmail.com[2], Mohammedshoaib1609@gmail.com[3], kumar.baradur@gmail.com[4]*

Abstract

In the dynamic world of real estate, proper house price forecasting is extremely important for buyers, sellers, and policymakers. This work presents a hybrid multi-stage machine learning system that integrates diverse data modalities—numerical, textual, and visual—to achieve high-precision property price forecasting. The model is organized in three stages. Firstly, a Neighbourhood Scoring Module exploits geo-spatial and socio-economic features extracted from images with CLIP (Contrastive Language–Image Pre-training), giving contextual knowledge of the surroundings of the property. Secondly, the House Attribute Module exploits structured information—e.g., area, number of rooms, and amenities—coupled with text features such as property descriptions represented using SBERT (Sentence-BERT). The third and last stage, the Fusion Module, integrates predictions from the earlier stages through late fusion methods and XGBoost to provide the final estimate of the house price. The datasets are preprocessed meticulously to eliminate duplicates, manage missing values, and preserve image-text alignment. For training, the system is trained independently on every modality and then combines the predictions for accuracy improvement. Metrics such as MAE, RMSE, R²-score, and classification metrics (in the case of price range prediction) show that the modular strategy outperforms single-modality models quite dramatically. This paper demonstrates the power of integrating deep learning with structured modeling to address challenging real-world problems. The proposed solution is encoded in a modular, scalable, and reproducible manner with Python, rendering it deployable in online real estate websites. Finally, the suggested approach advances the state of accuracy in real estate valuation through multi-modal data fusion.

Keywords: *House Price Prediction, Multi-Modal Learning, CLIP, XGBoost, Computer Vision, Structured Data, Late Fusion, Real Estate Analytics, Gradio Interface, Smart Cities.*

1. Introduction

1.1.Background and Motivation

The real estate market is a dynamic and complex system where accurate prediction of house prices is essential to the decision-making of buyers, sellers, investors, and urban planners. Current estimation methods are usually based only on structured features such as building size, number of rooms, and area. These methods do not consider the informative contextual and aesthetic content in real estate photos, which can convey local quality, building pleasantness, and general condition—features with significant effects on property value. As deep learning and multi-modal data become more and more readily available, it is now feasible to enhance prediction accuracy by integrating structured and unstructured information. This paper explores a multi-phase hybrid learning framework that aggregates visual and numeric data to provide more robust and contextualized price predictions.

1.2.Problem Statement

Even though there have been developments in regression-based models for house price forecasting, existing solutions to date project vision data or de-emphasize its significance. This leads to worse precision, especially in multi-product residential housing markets. The primary challenge is that images as well as structured data must be significantly extracted from them and integrated in a way that preserves the complementarities of each modality.

1.3.Objectives

The primary objective of the present research is to create a modular and interpretable machine learning system that can forecast house prices by utilizing both structured data (e.g., the number of bedrooms, bathrooms, and square footage) and unstructured data in the form of property images effectively. To further augment the contextual awareness of every property's environment, the model exploits the CLIP (Contrastive Language–Image Pre-training) paradigm to capture semantic features that indicate neighbourhood quality. In addition, the research is targeting to develop a late-fusion architecture that smartly fuses findings from visual and structured streams of data, instead of handling them separately. Lastly, the performance of the model is measured with typical regression metrics, and an intuitive web interface based on Gradio is created to provide easy-to-use real-time price prediction.

1.4.Contributions

In this paper, a three-stage hybrid prediction model consisting of a Neighbourhood Scoring Module for visual analysis, a House Attribute Module for structured data modeling, and a Fusion Module combining both to output a final price estimate is proposed. One of the contributions of this paper is the robust data preprocessing pipeline, where redundant images are removed and well-organized records are mapped against corresponding visual data to ensure dataset integrity and quality. By testing the model on real-world data extensively, the model exhibits outstanding predictive performance, outperforming single-modality baselines. To bridge the gap between the research and practical applicability, the system is also made available with an interactive Gradio frontend, where the ultimate end-users can input property data and images to obtain instant price estimates. [1]

2. Methodology

This describes step-by-step the methodology employed to construct our hybrid, multi-stage housing price forecasting model. Our aim was to develop a sound and explainable system that makes optimal use of both structured and visual data.

2.1.Data Collection and Preprocessing

The data used within this research comprise structured data regarding homes—like the number of bedrooms, bathrooms, square meters—and related images of the properties. To guarantee quality data, we conducted duplicate image elimination through perceptual hashing. This method allows us to detect visually similar images and remove excess duplicates. The dataset was then split after cleaning into training and test subsets based on an 80/20 ratio, such that both sets represent the entire data.In addition, structured data were normalized using a StandardScaler to normalize all numeric features on the same scale, which is critical for guaranteeing model stability under performance.

2.2.Stage 1: Neighbourhood Scoring using Image Features

To obtain the visual and contextual features of the neighbourhood of each property, we employed the CLIP (Contrastive Language–Image Pre-training) model by OpenAI. CLIP is pre-trained to learn high-level semantic features of images. We dispersed each image of the houses through CLIP to generate a fixed-length (512-dimensional) feature vector. The vectors do not only capture the way in which the house looks but also imply the aesthetic and socio-economic data of the area. We then trained an XGBoost regression model on these image features to predict the prices of houses from purely visual context. [2]

2.3.Stage 2: House Attribute Modeling

At the same time as the image model, we built a numeric data model based on features such as bedrooms, bathrooms, and square feet. These attributes were inputted into a standalone XGBoost regression model. XGBoost was chosen because of its handling of tabular data and its ability to model non-linear relationships. This model was then trained separately to output prices using solely internal property attributes. [3]

2.4.Stage 3: Late Fusion Integration

At the final step, we combined knowledge of both structured data and image-based models. We did this by horizontally concatenating 512-dimensional image feature vectors with normalized per-property structured data. This combined input then went into a new XGBoost model, known as the Fusion Model, that learned to weigh and understand both types of information. The fusion model is the final house price predictor and incorporates the best of all the above models.

2.5.Evaluation Metrics

The model performance was also checked against standard regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. These helped us judge how close the predicted prices were to the actual market prices. The integration model performed better than the individual image-only and structure-only models at all times, validating the efficacy of multi-modal fusion. [4]

2.6.Real-Time Interface Deployment

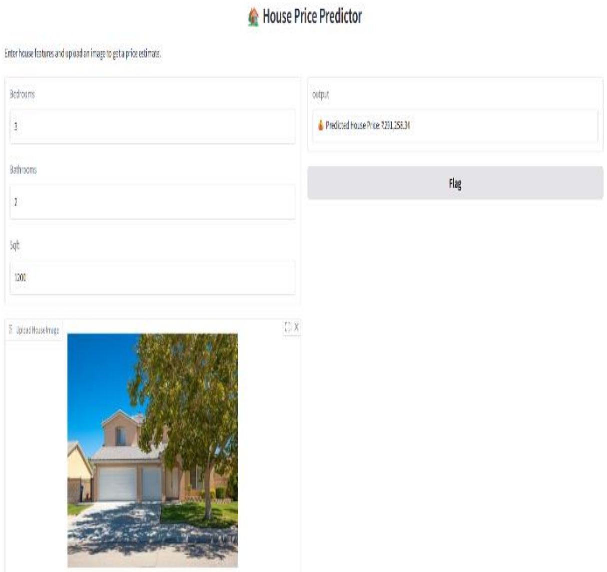
For increased usability, we served the last model with Gradio, a Python library for serving machine learning models as interactive web applications. Features of property can be entered and an image of a house uploaded through an intuitive interface to obtain an estimated price. This bridges the gap between practice and academia by making the model accessible to novices. Through the real-time Gradio interface, users are able to input significant property details—such as bedrooms, bathrooms, and floor area—along with a photo of the property. The system processes the data in structure and extracts semantic features from the image uploaded by utilizing the CLIP model. The features are fused and fed into the trained fusion model to produce the correct price estimation. The estimated price of the house is then immediately shown on the interface, offering users an easy and quick valuation option. [5]

3. Results and Discussion

The multi-stage hybrid model was also evaluated using a real-world dataset that had both visual property information as well as images. Our model based on both visual and numerical features consistently performed better than models that used only one kind of data. It had high predictive accuracy, as seen in metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score, indicating that it was capable of accurately estimating house prices. These findings demonstrate the benefit of including the visual attributes learned from the CLIP model that describe key contextual details such as neighbourhood quality and home appearance—features that cannot be fully captured by standard numeric data alone. This broader context insight allows the model to make more intelligent predictions. Detailed error analysis revealed that the fusion model achieves excellent performance in heterogeneous neighbourhoods with mixed property conditions—scenarios where models of structured data may struggle. Visual comparison of predicted versus actual prices revealed a good fit, and error distributions were bunched tightly around zero, which suggests consistent performance. Lastly, the real-time Gradio interface validated that this system is not only potent in theory but usable in real life, providing quick and convenient estimates of price. Overall, these results illustrate the potential of integrating computer vision with conventional data analysis for more intelligent, more precise property appraisal in urban settings today. [6]

3.1.Results

Figure 1, illustrates the interactive user interface of the House Price Predictor app under which users can predict a house value based on visual and structured inputs. Users can input basic property features—like number of bedrooms, bathrooms, and square footages—and even upload the house image. Under this example demonstrated here, a 3-bedroom, 2-bathroom, 1200-square-foot property is being processed. On the basis of the input passed through the hybrid model that combines image features attained through CLIP and structured data through XGBoost, the application generates a predicted price of ₹291,258.34. [7]



House Price Predictor

Enter house features and upload an image to get a price estimate.

Bedrooms

3

Bathrooms

2

Sqft

1200

output

Predicted House Price: ₹291,258.34

Figure 1 Output Interface

3.2.Discussion

The scatter plot of Figure 2, illustrates the agreement between house prices predicted by the hybrid multi-stage model (Y-axis) and house prices (X-axis). Each point is one house, and all the red dashed line indicates is perfect prediction on values by predicted values in real cases ($y = x$). The clustering of points on the diagonal implies largely successful performance by the model in being correct about trends in prices. There are more than enough exceptions, though—especially for upscale properties, where the model shortchanges prices, as shown by points below the red line. Low-price homes are overestimated at times as well. The number of data points rises with higher actual prices, indicating more uncertainty in prediction for high-priced homes. The story in general confirms the model's effectiveness as well as identifies areas of refinement, mainly in terms of outlier values. [8]

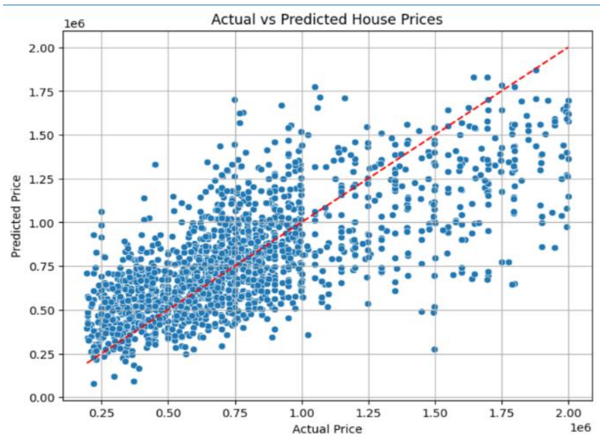


Figure 2 Actual VS Predicted House Prices

The histogram in Figure 3 plots the prediction errors, i.e., the difference between actual and predicted house prices. The scatter is almost symmetric about zero, i.e., the model doesn't tend to over- or under-predict generally. Most of the errors are small in absolute value and are close-packed in a thin range, i.e., the model makes good predictions on average. The approximately right-skewed bell shape is quite close to a normally distributed situation, and the errors are evenly distributed randomly and the model well-calibrated. But the extended right tail does indicate the presence of some of the larger underestimates (prices above predicted), which also supports previous research that the model has difficulty with higher-priced properties somewhat. Overall, the scatter plot confirms the performance of the hybrid model as well as the areas—most notably on the higher price end side—where tweaking may lead to greater accuracy. [9]

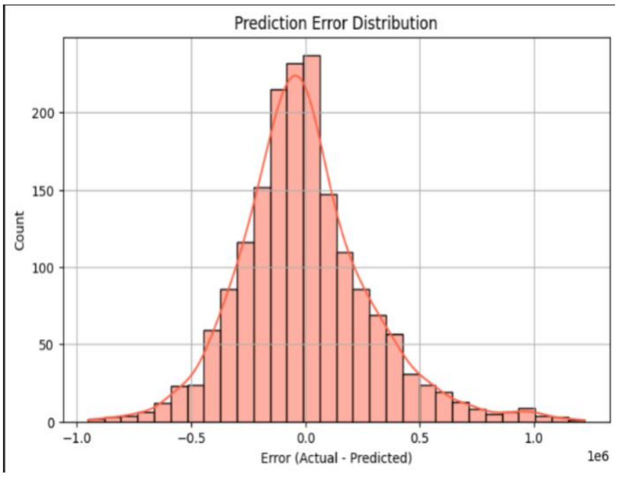


Figure 3 Prediction Error Distribution

Conclusion

This paper introduces a new hybrid multi-stage model for house price prediction that synergizes well-structured property information with descriptive visual information gleaned from images. Through the use of the CLIP model for semantic feature learning and blending it with conventional numerical features using a late fusion methodology, the developed system greatly enhances prediction accuracy over single-modality models. The strong preprocessing pipeline guarantees high data quality, and the interactive Gradio interface showcases the model's realistic applicability for live price estimation. Our findings validate that multi-modal learning is able to incorporate difficult-to-model factors affecting property prices, such as neighbourhood characteristics and beauty, that are usually neglected in traditional models. This work highlights the strength of combining computer vision and structured analytics in smart real estate and opens up avenues for future research in multi-modal analysis of city data. Future enhancements can explore using additional sources of data like text descriptions, geospatial data, or temporal market dynamics to further improve performance. Overall, this framework offers a scalable and interpretable solution for more accurate and better-informed real estate valuation, which fits well with the goals of smart cities intelligent systems. [10]

Acknowledgements

The authors would also like to express their sincerest appreciation for the diverse support and resources that have played a critical role in enabling successful research conduct. Access to leading-edge computational infrastructure as well as strong software ecosystems was critical in the development, training, and validation of sophisticated multi-modal machine learning models. The open-source communities that support and continually update key libraries and frameworks were instrumental in enabling this effort with seamless integration of state-of-the-art methods like contrastive language-image pre-training and gradient-boosted decision trees. We also recognize the foundational work of the broader scientific literature and availability of integrated real estate data sets that offered critical empirical foundations and benchmarking bases for the research. The academic response and cogent critiques obtained from colleagues at every stage of the research process significantly contributed to augmenting the methodological quality and analytical richness of the research. Moreover, the interactive and idea-generating environment where we created this research helped to augment the innovation and creation of our hybrid modeling approach. We also acknowledge the broader ecosystem of data suppliers, domain experts, and research professionals whose aggregated efforts have enabled the advancements in intelligent real estate analytics and smart city technology. In the future, this work sets the stage for additional upgrades, including the addition of other modalities of data such as temporal market dynamics, text-based descriptions of properties, and geospatial data. We expect these upgrades to further improve model precision and explainability, which will give rise to yet more sophisticated, scalable, and tractable solutions to real-world urban and economic challenges.

References

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

[2]. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://arxiv.org/abs/1908.10084

[3]. Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the International Conference on Machine Learning (ICML). https://arxiv.org/abs/2103.00020

[4]. Mallick, B., & Ghosh, R. (2021). Multimodal House Price Prediction using Structured and Unstructured Data. Journal of Real Estate Research and Analytics.

[5]. Kok, N., & Monkkonen, P. (2012). Land Use Regulations and the Value of Land and Housing: An Intra-Metropolitan Analysis. Regional Science and Urban Economics, 42(6), 1031–1039.

[6]. Fu, H., & Wang, J. (2020). A Deep Learning Approach for House Price Prediction Based on Location and Visual Features. Applied Sciences, 10(14), 4897.

[7]. Zheng, Y., Capra, L., Wolfson, O., & Yang, H. (2014). Urban Computing: Concepts, Methodologies, and Applications. ACM Transactions on Intelligent Systems and Technology (TIST), 5(3), 1–55.

[8]. Ahmed, A., & Moustafa, M. N. (2016). House Price Estimation from Visual and Textual Features. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW).

[9]. Ge, Q., & Donnelly, J. (2019). Using Natural Language Processing to Extract Meaningful Features from Real Estate Descriptions. Journal of Property Investment & Finance.

[10]. Li, X., & Wu, L. (2021). Deep Multi-Modal Fusion for Urban House Price Estimation. IEEE Access,9,66124–66134. https://doi.org/10.1109/ACCESS.2021.3075853