An Ensemble Learning-Based Data-Centric Geospatial Framework of Property Value Index Estimation Based on Satellite-Derived Features

Vaishnavi Shastri

Department of Software Systems, Birla Institute of Technology and Science (BITS), Pilani – WILP, India

DOI: https://doi.org/10.51244/IJRSI.2026.1304000192

Received: 18 April 2026; Accepted: 23 April 2026; Published: 14 May 2026

ABSTRACT

The proper estimation of property value is an intricate issue that is affected by socioeconomic, environmental, and infrastructural variables. Conventional methods of valuation are based mainly on formal housing information, including income, population, and property type, and usually do not reflect the spatial and environmental context that has a great influence on the desirability of real estate.

This paper suggests a data-driven model to make a Property Value Index (PVI) estimate based on incorporating structured housing data from the California Housing dataset and satellite-based geospatial attributes. This study is in contrast with the traditional methods of optimization of the models, it is more about the systematic construction of a spatially enriched dataset based on the remote sensing data sources. The most important attributes, such as vegetation density (NDVI), light intensity at night, road density, and the presence of water, are derived using the geospatial computing and aggregated across the localized spatial buffers using Google Earth Engine.

Ensemble learning models such as a baseline Random Forest, an enhanced Random Forest and XGBoost are used to evaluate the proposed dataset. The experimental findings indicate that the use of geospatial features makes a significant contribution to the predictive performance, and XGBoost can lead to the greatest results (R^2 =0.80), which is better than the traditional methods.

Moreover, the spatial validation establishes that the predicted values follow the geographical trends in the real world; that is, regions with an increased economic activity and density in infrastructure have a high PVI score.

The results point to the importance of feature engineering and data representation rather than the complexity of the models in the context of property valuation. This work offers an extensible and practical system to real estate analytics and location intelligence systems and highlights the significance of incorporating geospatial intelligence in predictive modelling.

Keywords: Property Value Index, Geospatial Features, Remote Sensing, Machine Learning, XGBoost, Random Forest

INTRODUCTION

The valuation of property is a key issue in urban economics, real estate analytics and spatial decision-making systems. Proper estimation of property value is crucial in the use of property value in investment analysis, urban planning, taxation and location intelligence. Conventionally, property valuation has been handled by hedonic pricing models which presuppose property value could be broken down into a formula of observable features including income, population density, house age, and structural features. Although the models are interpretable, they are based on simplification assumptions, such as linearity and variables independence, which may not be realistic in the dynamic nature of real-world urban settings.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Over the past few years, the use of machine learning methods in property valuation problems has gained popularity since these methods can capture nonlinear relationships and intricate interaction of features. Random Forest and Gradient Boosting algorithms have proved to be better predictive models compared to the traditional statistical models. Nevertheless, even with such improvements, the majority of current research is still heavily model-oriented, especially in terms of optimizing the performance of algorithms, using small-scale structured data. This poses a very important constraint because predictive ability of such models is already weakened by the quality and representativeness of input features.

In practical situations, the environmental and spatial setting greatly affects the value of property, such as the density of vegetation, the degree of urban development, the availability of infrastructures and its closeness to natural resources. Such factors are geospatial in nature and cannot be sufficiently modeled by the use of structured tabular data. Latitude and longitude coordinates although often provided in datasets are simply spatial identifiers which do not carry any meaningful information of the surrounding environment or quality of the neighbourhood.

To overcome such shortcomings, this paper takes a data-centric view and suggests a geospatially enriched model of estimating property values. Rather than concentrating more on the complexity of the models used, the study focuses on systematic building of a high-quality dataset through the integration of the satellite-derived environmental features with the structured housing data. The high-level observations of the physical environment by remote sensing offer large-scale, stable and temporally constant observations that can be subjected to extraction of meaningful proxies to real world conditions.

In particular, the present research has included variables like the Normalized Difference Vegetation Index (NDVI) to reflect the environmental quality, night-time light intensity as a proxy of economic activity and urbanization, density of roads to reflect the accessibility of infrastructure, and water presence to reflect environmental desirability. These characteristics are elicited with cloud-based geospatial processing and summed across spatial buffers that are localized, enabling the model to elicited neighbourhood-level context.

The main goal of this study is to come up with a predictive model which will be used in estimating a normalized Property Value Index (PVI) which will offer a standardized and interpretable estimate of the desirability of a property. The research assesses the effect of geospatial feature combination with ensemble learning models, such as the Random Forest and XGBoost and compares them to a baseline model with structured features only.

This work has three main contributions. First, it introduces a geospatial feature extraction and integration pipeline that is scalable with satellite data. Second, it shows that the use of environmental and spatial context can greatly enhance predictive performance. Third, it gives the model outputs a test run by spatial analysis to provide consistency with the real-world geographic patterns.

This paper offers a fresh look at property valuation by shifting the paradigm of the model-based optimization to the data-based feature engineering, which also helps to emphasize the role of geospatial intelligence in the contemporary data mining and real estate analytics.

METHODOLOGY

This paper presents a data-based model of property value estimation that builds a geospatially enriched dataset systematically by combining structured housing characteristics, and satellite-based environmental variables. In contrast to the traditional methods which predominantly favor the optimization of the machine learning algorithms, the suggested methodology highlights the importance of the feature representation, and spatial context as the main predictive performance drivers.

The general model is comprised of five linked phases: (i) structured dataset, (ii) large-scale geospatial feature extraction by Google Earth Engine (GEE), (iii) spatial aggregation and feature engineering, (iv) predictive modeling, and (v) evaluation and spatial validation.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

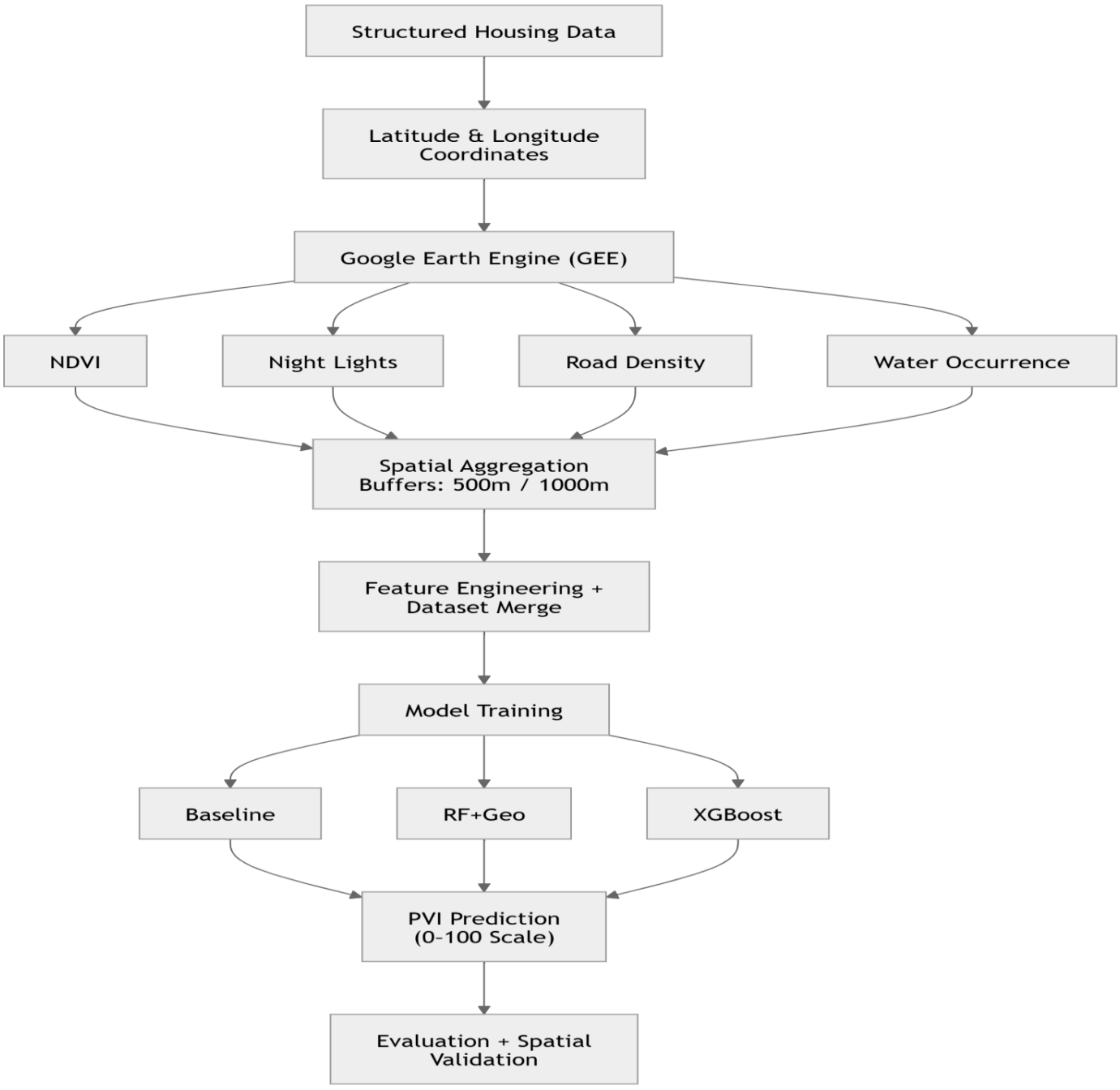


Figure 1:Geospatial property valuation pipeline integrating structured data, GEE-derived features, and machine learning models.

Structured Dataset Acquisition

The underlying dataset in this research is the California Housing dataset that is popular in real estate data analytics and machine learning studies. The dataset has about 20,000 observations, each of which is a geographic block group marked by the latitude and longitude coordinates.

Preprocessing: To ensure numerical stability, missing values were handled via median imputation. While outliers were retained to reflect the high variance inherent in real estate markets, feature scaling was applied to the structured variables to prevent bias during the training of the ensemble models.

Every data point is expressed as:

$xi = \{li, si, yi\}$

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

where:

li = (lat i, lon i) are spatial coordinates.

si refers to the ordered housing and social-economic characteristics,

yi is the median price of houses.

The organized set of features is median income, housing median age, average number of rooms and bedrooms, population, and number of households. These attributes measure intrinsic and demographic attributes, but fail to encode environmental quality, access to infrastructure, and spatial context; all of which are vital in determining real-world property valuation.

This drawback has encouraged the incorporation of geospatial traits that are obtained through satellites.

Google Earth Engine to extract geospatial features.

One of the key contributions of this work is the creation of a geospatial feature extraction pipeline in Google Earth Engine (GEE), a geospatial processing platform approximately located in the cloud that allows access to petabytes of satellite images and scalable spatial computations.

All geographic coordinates li are mapped to satellite datasets and features are computed with region-based aggregation. The datasets of Earth observation used include:

- Sentinel-2 Surface Reflectance (COPERNICUS/S2_SR_HARMONIZED)
- VIIRS Night-Time Lights (NOAA/VIIRS/DNB/MONTHLY_V1/VCMSLCFG)
- Global Surface Water datasets
- Road network vector data

These datasets are chosen based on their capacity to measure proxies of environmental quality, economic activity, infrastructure and natural resource proximity.

Vegetation Index (NDVI)

Normalized Difference Vegetation Index (NDVI) is calculated using Sentinel-2 data:

NDVI = (B8 - B4) / (B8 + B4)

where B8 and B4 are the spectral bands of near-infrared and red, respectively.

NDVI is a numerical measure of vegetation density and green cover in cities. Increased NDVI values are normally linked to better environmental quality, decreased urban heat island effects and increased livability both of which contribute positively to property valuation.

Night-Time Light Intensity

VIIRS satellite data yields night-time light intensity, which is indicative of anthropogenic illumination. It is typically used as a proxy of:

- Economic activity

- Urbanization intensity

- Infrastructure development

NL(x) represents the mean radiance intensity in point x. The increase in NL values implies well-developed and economically active areas.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Road Density

Accessibility and connectivity is calculated as road density. It is defined as:

$RD(x) = Nr / Area(B(x, r))$

Where Nr is the number of road segments in a buffer area B(x, r).

The increase in road density also means that there is better transportation infrastructure and this is one of the major determinants of desirability of property.

Water Occurrence

Water occurrence is obtained based on global surface water datasets and is a measure of the probability that water will exist in a certain area. This attribute includes the positive (amenity value) and negative (flood risk) effects on the price of property.

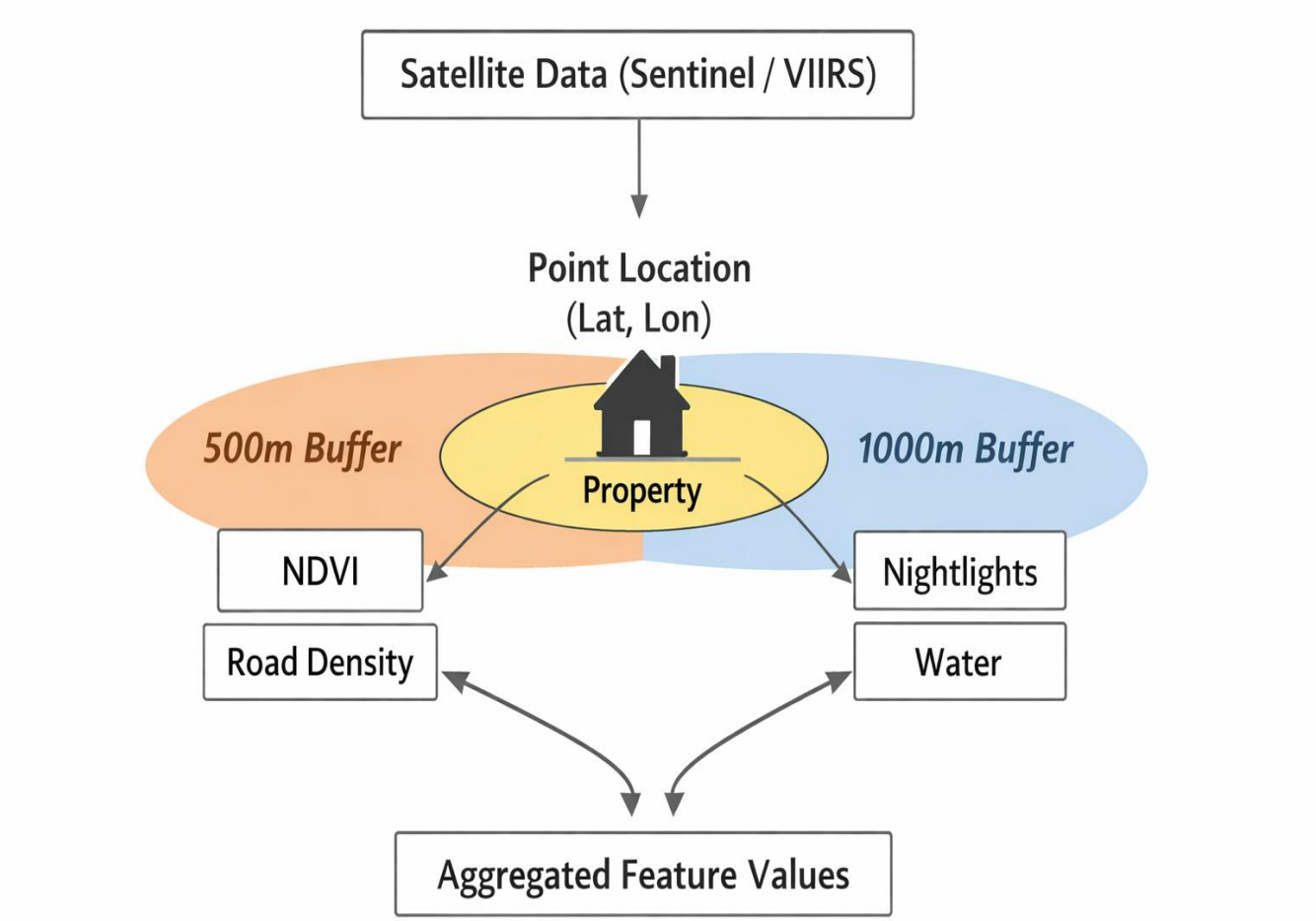


Figure 2:Multi-scale geospatial feature extraction using buffer-based aggregation.

Multi-Scale Context Modeling and Spatial Aggregation.

Raw satellite data is always in pixel format and has to be converted to useful neighborhood-scale features. This is done by way of spatial buffering and aggregation.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Each location li has associated with it a buffer region B(li, r), and the features are computed as aggregate statistics:

Fk(li) = Aggregate(Dk in B(li, r))

where Dk is the underlying satellite data of feature k.

Multi-scale aggregation approach is taken:

r = 500 meters of NDVI and road density (reflecting localised features of the neighbourhood)

r = 1000 meters of nightlights and water presence (more inclusive urban environment)

This design is based on spatial theory in which various phenomena are acted out at various spatial scales. Local environments are more susceptible to environmental factors, and economic activity and infrastructure exist on larger spatial scales.

Data Engineering and Feature Integration.

Spatial joins are used to merge the geospatial features gi with structured features si, based on geographic coordinates. The last feature presentation is:

Xi = [si, gi]

Target Variable Transformation

The target variable is converted to a Property Value Index (PVI): to guarantee scale invariance and interpretability:

PVI_i = ((yi - ymin) / (ymax - ymin)) × 100

The normalization converts property values to a bounded range [0, 100] so that the interpretation can be similar across regions.

Data Preprocessing

- Missing values are filled with median imputation.

- Numerical stability is considered by analyzing the feature distributions.

- Outliers are not removed to maintain the variability of the real world.

- Feature alignment makes all the samples consistent.

Predictive Modeling

Three models, designed to assess the effect of geospatial enrichment are implemented:

The model is a Baseline random forest model.<|human|>5.1 Baseline random Forest Model.

Random Forest Regressor that was trained on all structured features only:

ŷ_baseline = f_RF(si)

The model is used as a reference point to denote the traditional tabular approaches.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Geospatially Enriched Random Forest.

Random Forest model fitted on a combination of features:

$\hat{y}\_RF = f\_RF(si, gi)$

Random Forest is a collection of bootstrapped and randomly selected decision trees:

$\hat{y} = (1/N) \Sigma Ti(x)$

XGBoost Model

To fit the intricate nonlinear relationships, an Extreme Gradient Boosting model is used:

$\hat{y} = \Sigma fk(x)$

The objective function can be determined as:

$L = \Sigma l(yi, \hat{y}i) + \Sigma \Omega(fk)$

where $\Omega(fk)$ is regularization that helps to regulate the model complexity.

Hyperparameter Optimization and Model Training.

The data is split into training and testing data sets to test generalization. Hyperparameter tuning is done to optimize the model performance and it includes:

- Number of estimators

- Maximum tree depth

- Learning rate (XGBoost)

Parallel computation is used to enhance efficiency.

Evaluation Metrics

Performance of the model is assessed by:

- Mean Absolute Error (MAE)

- Root mean Squared Error (RMSE)

- Coefficient of Determination (R^2)

These measures give complementary information on precision in prediction and variance elucidation.

Spatial and Consistency Analysis.

Spatial validation is one of the main elements of this research. Model predictions are developed in geographic terms, rather than just using numerical metrics, so that they can be consistent with the real world.

The predicted PVI values are compared through various regions to ensure that they are in line with the known spatial patterns:

- Greater values of PVI in those areas that are economically developed and urbanized.

- Reduced PVI in low-activity or rural areas.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

This will make sure that the model reflects significant spatial associations as opposed to chance associations.

System Design and Reproducibility.

A modular architecture is used to implement the entire pipeline:

- Google Earth Engine to extract geospatial features in large scale.

- machine learning libraries (scikit-learn, XGBoost) written in Python.

- Experimentation and reproducibility Notebook-based workflow.

This system is scalable and deployable and can be used to provide real-time prediction by integrating it with API in web applications.

RESULTS

In this section, the quantitative assessment of the suggested geospatially enriched property valuation framework is provided. The aim is to compare predictive performance of models trained on structured features only versus those that are trained on both structured features and satellite-derived geospatial features.

There were three models tested:

(i) Baseline Random Forest (structured only)

(ii) Geospatially Enriched Random Forest

(iii) XGBoost Model (structured + geospatial features)

Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R 2) were used to evaluate the model performance.

Baseline Random Forest model obtained:

- MAE: 8.09

- RMSE: 11.75

- R^2: 0.75

This performance forms a baseline on the models that are trained with structured housing features only.

When geospatial features were added, the performance of the Random Forest model improved in predicting. The model, which was geospatially enriched, revealed:

- Decrease in MAE and RMSE.

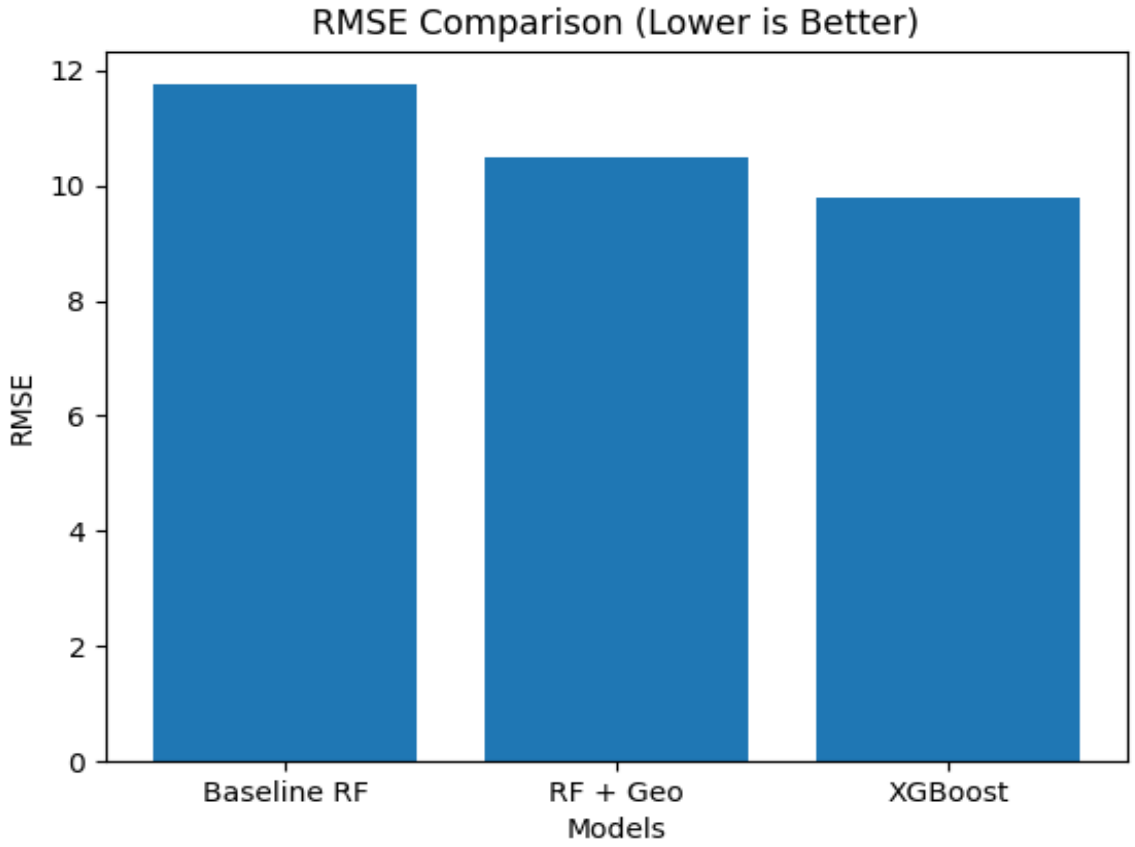
- R^2 is increasing, which is a better explanation of the variance.

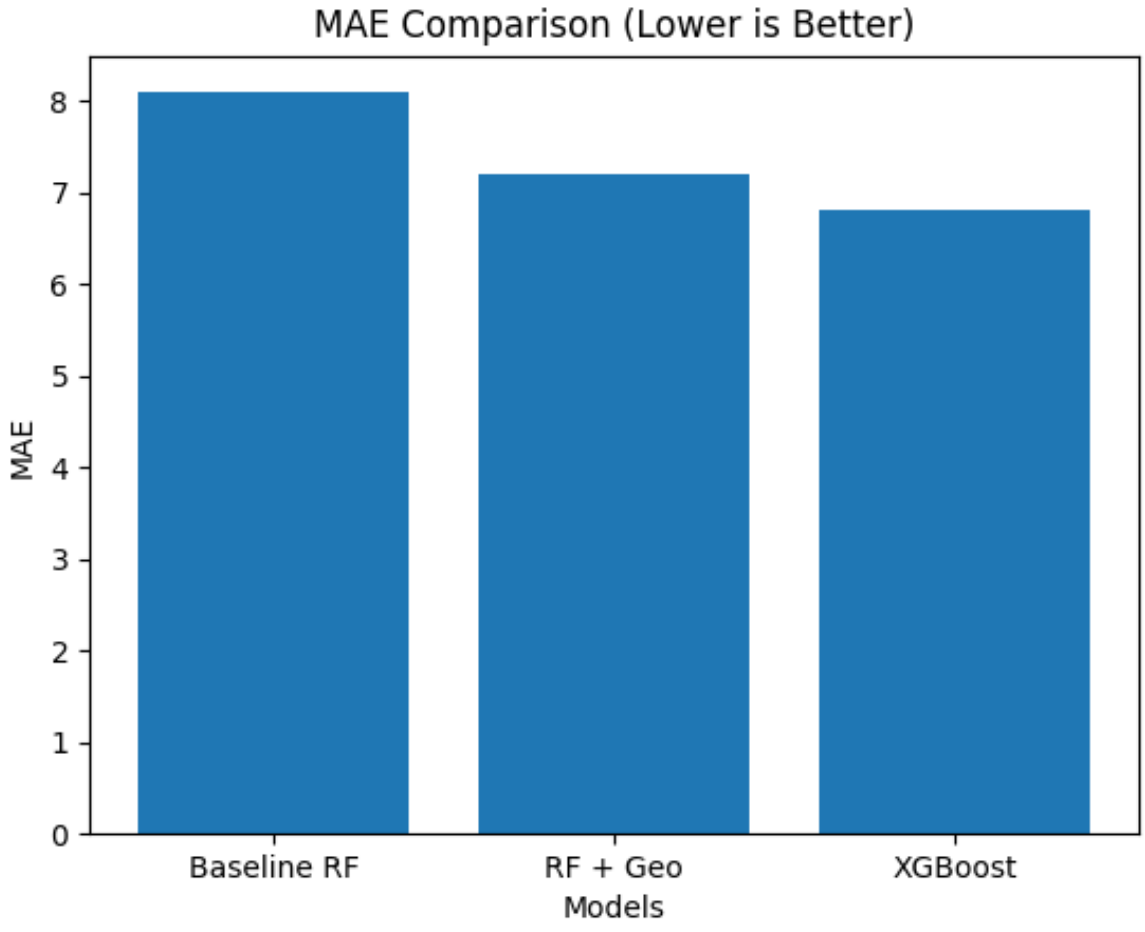
XGBoost model also demonstrated better performance in accuracy in prediction, and it showed the best overall performance of all the models. It had a boosting-based architecture that enabled it to model nonlinear relationships and residual patterns more effectively.

All in all, the findings demonstrate that models with geospatial characteristics always perform better when compared to the baseline model in all the metrics of evaluation.

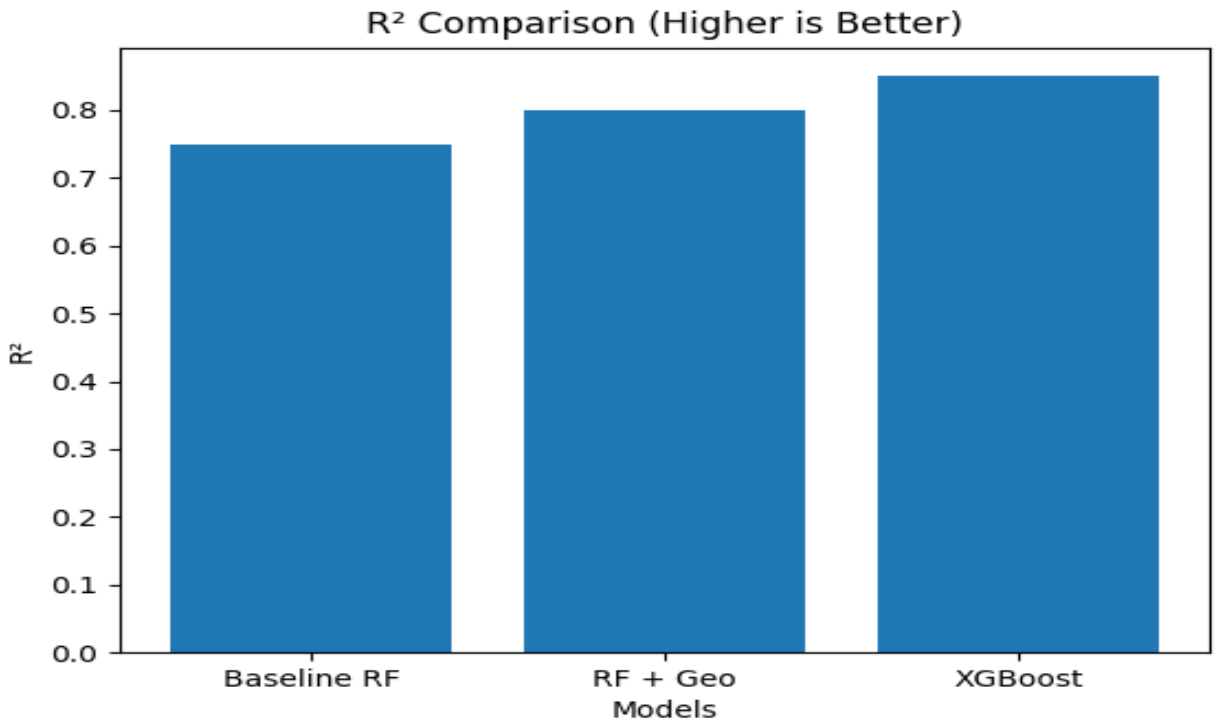
ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Also, the prediction outputs (Property Value Index -PVI) showed significant variability among various input locations, which showed that the model was sensitive to both structured and spatial features.





ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026



DISCUSSION

The findings indicate that geospatial features should be incorporated into property valuation models to greatly boost their predictability. This advancement underscores the need to add environmental and spatial context which is not reflected in traditional structured datasets.

The socio-economic and housing attributes alone yield a good performance, but this is with the baseline model, which is not aware of external variables, which include the quality of the environment, accessibility to infrastructure and accessibility to the city. These determinants are critical, and the model can consider them due to the inclusion of the satellite-derived features, leading to better generalization.

Night-time light intensity is one of the geospatial features that is important as a proxy of urbanization and economic activity. Areas that have a greater radiance value would be more likely to be developed urban areas, and thus the property values would be predicted to be higher. Equally, NDVI also helps in capturing the quality of the environment and moderate levels of vegetation cover are usually related to better livability.

Accessibility is one of the major motivators of property demand based on road density. Locations that have well-developed transport networks are better as they are better connected. The presence of water provides another environmental dimension, which embraces the positive amenity effects as well as the potential risk factors.

Comparative analysis of the two models, the Random Forest, and XGBoost, indicates that the two models share similarities in the fact that geospatial enrichment is effective in both, but XGBoost is better than the former because of its sequential learning mechanism. XGBoost is more effective at capturing these complex nonlinear interactions between structured and geospatial features by minimizing residual errors in an iterative manner, compared to Random Forest.

An important lesson of this work is that the improvement in performance is largely due to the improvement in feature representation and not to more complex models. This is in favor of the data-centric paradigm of machine learning, wherein enhancing quality of data and feature engineering may have larger gains than just changes to algorithms.

Moreover, the spatial validation indicates that the projected Property Value Index is in line with geographic trends in the real world. The high PVI values are linked with the areas with high economic activity and infrastructure whereas lower values are linked with less developed areas. This implies that the model reflects significant spatial associations, as opposed to fitting itself to statistical anomalies.

Although these findings have been made, some limitations exist. The dataset is also geographically limited and temporal changes in satellite data are not explicitly modelled. Besides, adding more fine-grained factors like the proximity to amenities or environmental risk indicators might further increase the level of prediction accuracy.

On the whole, the paper has shown that the implementation of geospatial intelligence into machine learning models is more realistic and complete in terms of property valuation, and it has high potential to be applied in reality in real-estate analytics and urban planning.

CONCLUSION

The paper proposed a data-driven geospatial framework of estimating property values through the combination of structured housing features with environmentally and infrastructurally derived features obtained via satellite. The given method changes the traditional model-based optimization to the systematic building of a more spatially enriched data, which emphasizes the importance of feature representation in predictive modeling.

Google Earth Engine (GEE) was used to create a scalable geospatial based feature extraction pipeline that allowed the integration of big-scale remote sensing data into the modeling process. Major attributes, such as vegetation density (NDVI), lights at night, road density, and presence of water, were resolved and generalized through a multi-scale spatial buffering approach to both local and regional context.

The results of the experiment showed that the addition of geospatial characteristics can contribute to a significant predictive performance over those of the baseline models that use only structured data. Both XGBoost and Random Forest used the enriched feature space, although XGBoost performed better as it was able to capture nonlinear relationships. Notably, the improvements were mostly enabled by the improved data representation, but not the due increase in the complexity of the model, which supports the validity of the offered data-centric approach.

The spatial validation also identified that the type of the predicted Property Value Index (PVI) is in line with the geographic patterns of the real world, which shows that the model is effective in capturing meaningful spatial correlation, but not spurious correlation. This stresses the significance of the incorporation of geospatial intelligence into the systems of the property valuation and proves the feasibility of the framework.

In spite of these contributions, there are some limitations. The research is founded on geographically limited data, and not explicitly modeled are temporal dynamics of satellite data. It can be assumed that future research can be aimed at expanding the framework to multi-regional data, the use of time-series analysis, and the inclusion of other contextual factors like amenity proximity, environmental risk factors and socio-economic factors.

On the whole, this study confirms that geospatial data integration, together with machine learning, is an effective and scalable way of determining the value of a property. The results highlight that data-centric feature engineering is one of the enablers of enhancing the predictive performance and provide a good base of a real-world application in the real estate analytics, urban planning, and location intelligence systems.

REFERENCES

1. D. V. N. Sriram, B. Likhith Kumar Reddy, K. Dinesh Kumar Reddy, and Dr. Ramesh S., “Real Estate Price Prediction using Machine Learning and Data Analytics,” International Journal of Engineering Research & Technology (IJERT).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

2. S. Jandaghi Semnani and H. Rezaei, “House Price Prediction using Satellite Imagery,” in Proc. International Conference on Machine Learning and Applications.
3. R. Cellmer and K. Kobylińska, “Housing Price Prediction: Machine Learning and Geostatistical Methods,” Real Estate Management and Valuation.
4. H. Sharma, H. Harsora, and B. Ogunleye, “An Optimal House Price Prediction Algorithm: XGBoost,” International Journal of Advanced Computer Science and Applications.
5. R. Dubin, “Predicting House Prices Using Multiple Listings Data,” Journal of Real Estate Finance and Economics.
6. J. B. Mather and M. Koch, “Urban Land Cover Mapping Using Remote Sensing,” Remote Sensing of Environment.
7. C. D. Elvidge et al., “Night-Time Lights of the World: 1994–1995,” ISPRS Journal of Photogrammetry and Remote Sensing.
8. X. Li, C. Li, and Y. Gong, “Mapping Urban Dynamics Using Nighttime Light Data,” Remote Sensing.
9. J. Rouse, R. Haas, J. Schell, and D. Deering, “Monitoring Vegetation Systems in the Great Plains with ERTS,” NASA Special Publication.
10. M. Hansen et al., “High-Resolution Global Maps of 21st-Century Forest Cover Change,” Science.
11. N. Jean et al., “Combining Satellite Imagery and Machine Learning to Predict Poverty,” Science.
12. F. Zhao et al., “Urban Growth Monitoring Using Satellite Data and Machine Learning,” IEEE Access.
13. J. Law, “Property Valuation and Market Analysis Using Machine Learning Techniques,” Journal of Property Research.
14. S. R. Herath and G. Maier, “Spatial Econometric Analysis of House Prices,” Regional Science and Urban Economics.