International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 5, Sept-Oct-2024, ISSN (Online): 2395-566X

House Price Prediction Models with Noise-Injected Data Using Machine Learning

S.Shanmathi, V.Rajeswari, V.Chaitanya, T.Navya, P.Vasudeva Rao
Electronics and Communication Engineering
GMR Institute of Technology, Rajam

Abstract- Using machine learning techniques, notably linear regression, the project "House Price Prediction Models with Noise-Injected Data Using Machine Learning" aims to improve house price predictions. Data collection, preprocessing, and the incorporation of environmental elements like noise levels into the model are among the goals of the study. The study's data base consists of publicly available datasets from real estate sources and websites like Kaggle. To create a reliable prediction model, the methodology uses an organized procedure that includes data collection, preprocessing, feature engineering, exploratory analysis, model selection, and comparison analysis. Accurately predicting house prices is achieved by the use of linear regression, and the model's performance is assessed using metrics such as Mean Squared Error (MSE) and R-squared (R^2). The findings show that important variables like housing size, location, and noise levels have a big impact on the forecasts. High R-squared values and a low Mean Squared Error confirm the model's good predictive ability and validate that it is a reliable tool for projecting property prices.

Index Terms- Linear regression, Machine Learning, Model evaluation.

I. INTRODUCTION

Real-estate prices are influenced by a wide range of factors, including both property characteristics and noise level conditions. The impact of noise levels and other key factors is to create a thorough machine learning model that forecast home values by taking into both traditional property feature (such as location, size, number of bedrooms and amenities) and noise factors. In order to produce more precise and nuanced forecast that combines structured data (noise level) with unstructured data (property features) and uses linear regression techniques. Metrics like R2, mean absolute error (MAE), root mean squared error (RMSE) are used to choose the model. This project shows how to use data science and machine learning approaches to solve the real-world problems by providing a holistic view of factors affecting home prices and contributes to more informed decision making for buyers, sellers and real estate developers for House prediction. The "House Price Prediction" project is an applied data science endeavor that uses machine learning methods to forecast real estate values. In order to create a prediction model that can estimate home prices based on numerous features, the main goal of this research is to analyze historical real estate data. The dataset that was used contains details about the location, square footage, and number of bedrooms and bathrooms of a house. The creation of an accurate linear regression model for estimating home prices is the main goal. The interpretability and efficacy of linear regression in illuminating the correlations between the target variable (home price) and characteristics makes it the preferred method. Data collection, cleaning, and preprocessing are important phases in the project. Feature engineering is then used to improve the prediction ability of the model. To make sure the model is generalizable, its performance is evaluated using metrics like R-squared and cross-validated. The project intends to help real estate agents and potential purchasers make educated judgments in the real estate market by using a trustworthy prediction algorithm. Understanding price patterns and appraising properties according to their attributes can be aided by the insights this model generates. Property price forecasting has grown more difficult. Real estate pricing models have historically placed a strong emphasis on features like amenities, size, and location. These models, however, frequently fail to take into account the important influence that environmental conditions have on property values. A home's appeal and, hence, its price in a busy metropolitan setting are greatly influenced by elements like air quality, noise pollution, and accessibility to green areas. This uses a multi-modal method to combine noise levels and proximity to green spaces with standard property data. Through the acquisition of real-time noise data, the model provides a more precise and all-encompassing understanding of the variables influencing property pricing. More accurate forecasts may be made with this improved method, which benefits buyers, sellers, and real estate professionals who have to evaluate homes in terms of both structural and noise quality. A new method in real estate analytics, the incorporation of noise level parameters into home price forecast reflects the growing

© 2024 IJSRET

recognition of the significance of noise quality in urban living. The goal of this study is to shed light on how urban surroundings, particularly in fast developing cities like Bangalore, might influence residential decisions in addition to the monetary worth of real estate. The process of predicting home values is complex and requires examining a wide range of variables pertaining to a property's physical attributes as well as the surrounding area's noise pollution. Conventional models tend to highlight the property's interior attributes, like the age, square footage, and number of bedrooms. To improve the accuracy of these forecasts, there has been a renewed focus in recent years on utilizing outside variables like location data and ambient elements like noise levels. There is a direct correlation between noise pollution and living standards. Loud noises can make a place less acceptable overall, cause stress, and interfere with health. The proximity of highways, airports, or vibrant nightlife districts makes this feature particularly important. Property values tend to be higher for comparable properties located away from main sources of noise in quieter neighbourhoods. Many sources of noise data can now be gathered, including real-time applications, government noise maps, and reports from local communities. This is made feasible by technological improvements. Combining modern data on location and noise levels with traditional property attributes is crucial for developing a more accurate house price prediction model.

II. LITERATURE SURVEY

Separating up the body of research on house price prediction into categories based on machine learning and deep learning techniques. Numerous places, including the US, Europe, and developing nations like Brazil, have produced studies on property price prediction. This indicates that accurate housing price forecasting is of global interest. Utilizing a range of data sources, such as satellite photos and information on public buildings, helps to enhance prediction models. The study emphasizes how house price prediction algorithms are incorporating sophisticated machine learning techniques like ensemble approaches and neural networks. These methods have demonstrated promise in capturing intricate feature interactions, resulting in predictions that are more accurate. In order to improve prediction accuracy, the suggested model incorporates heterogeneous data in an attempt to close these gaps.

The article "Prediction of House Pricing Using Machine Learning with Python" examines key elements—such as size, age, type, location, and amenities—that have an impact on property pricing. It demonstrates how machine learning techniques may significantly improve prediction accuracy, particularly when they use regression methods and artificial neural networks (ANN). The poll also emphasizes the importance of graphical and numerical tools for data display, as well as the evolving role of housing as an investment vehicle. The social effects of house price plans, including affordable housing and urban growth, are also covered.

"House price prediction using machine learning" emphasizes how important it is to estimate home prices accurately in order to assist both buyers and sellers. It talks about the challenges real estate agents confront and stresses how important it is to develop reliable forecasting models in order to set reasonable pricing. outlines the procedures for collecting and pre-processing data, which are essential to ensuring the quality of the data before analysis. A comparison analysis is used to evaluate multiple machine learning models, and the model with the highest performance is selected. Linear regression is used in conjunction with the Python package scikit-learn because it can reveal connections between features and home prices. The study also highlights the use of real-world data to improve the model's relevance and efficiency, showing how machine learning may be applied in practical contexts.

With a focus on scene perceptions including views of the sky and greenery, this research review highlights how the Hedonic Price Model (HPM) can be used to investigate how environmental elements affect property values. Previous studies have looked into the relationship between housing costs and urban amenities, but this study closes a gap in the literature by include scene perception inside the HPM framework. The study emphasizes the importance of environmental factors in real estate value through the use of an open-access dataset by discovering that characteristics like greenery and sky openness significantly raise house prices in cities like Beijing and Shanghai.

This literature study emphasizes the importance of data quality, model selection, and the application of machine learning techniques in order to estimate Shenyang's pricing for used homes. The study concludes that XGBoost offers superior prediction accuracy when comparing the models of Lasso Regression, Ridge Regression, and XGBoost. It also highlights how data visualization can help real estate industry participants better understand complex market dynamics and provide them with relevant information. The paper provides a sound methodology for evaluating housing costs in urban areas, which advances the field of real estate economics.

Using satellite and street view photos to estimate home prices emphasizes the value of visual information in real estate appraisal. By capturing intricate spatial linkages and contextual information more accurately than conventional models, it presents a Multi-Head Gated Attention model that improves spatial interpolation techniques and dramatically increases the accuracy of house price projections. Furthermore, by using Geographically Weighted Regression (GWR) to examine the ways in which street-level perceived characteristics affect real estate values, it is possible to discern the varied effects of perceived beauty and livability on pricing

© 2024 IJSRET

in various metropolitan regions. Collectively, these studies highlight how combining sophisticated modeling methods with visual data can improve estimates of house prices, pointing to a move in real estate appraisal toward more sophisticated and data-driven methods.

III. METHODOLOGY

The house price prediction using machine learning, incorporating noise level information can enhance the model's accuracy and robustness. Here's a detailed methodology for integrating noise level data into the predictive model.

1. Data Collection

Gather noise level data from reliable sources, such as environmental monitoring agencies or urban studies. This data should include measurements of noise pollution in various areas where the housing data is collected. For your project "Smart Home Price Prediction: The Impact of Noise Levels and Other Key Factors in Bangalore," you can rely on a number of reliable sources to gather data on noise levels, ensuring thorough and accurate coverage. Governmental organisations that frequently monitor environmental conditions, including noise pollution, including the Central Pollution Control Board (CPCB) and the Karnataka State Pollution Control Board (KSPCB). These organisations also produce reports tailored to individual regions, like Bangalore. Furthermore, city-specific data may be provided via municipal and local government portals, or they may work with environmental organisations to track urban noise. Studies involving noise pollution are frequently conducted by academic institutions such as the Indian Institute of Science (IISc) in Bangalore and urban planning organisations; hence, their findings are invaluable resources.

2. Data Preprocessing

Making ensuring the data is clean, consistent, and prepared for integration with the rest of the housing data is essential for the pretreatment of the noise level data. In order to address any mistakes or inconsistencies in the dataset, you must first do data cleaning. To do this, outliers that might indicate unusual readings or mistakes made during the data collection process must be found and eliminated. Such outliers can be found using methods like standard deviation thresholds or box plots, which visualise the data distribution. Imputation methods such as mean, median, or mode imputation can be used to fill in the blanks for missing values. Since median imputation is less sensitive to outliers than mean imputation, it is frequently more robust depending on the distribution of the noise data.

3. Feature Engineering

Using noise level data for feature engineering, you can develop new features that more accurately reflect its influence on house values. Determine the average noise level in a specified radius surrounding each property first. This can be done by computing the average noise levels of all monitoring stations or data points within a specific distance (such as 500 meters or 1 kilometre) from each property. This can be done using spatial data techniques like distance-based aggregation. This component will assist the model comprehend how ambient noise varies with location by offering a localised picture of noise pollution. Decibel thresholds can also be used to group the noise data into bands or levels, such as low, medium, and high.

Finally, investigate correlations and interactions between noise levels and other dwelling characteristics (such as price, size, and neighbourhood quality) using visualisation tools like scatter matrix plots from the pandas library. You can see how the noise data connects to these factors and spot any patterns, like if higher noise levels are associated with smaller or cheaper homes, by using the scatter matrix. This can also assist you in determining which feature combinations would be beneficial for your prediction model or in detecting multicollinearity. You can provide the model more complex insights into how noise affects house prices by creating features connected to noise in various ways.

4. Model Selection

Random Forest is a solid option for your project's model selection stage because of its resilience in managing complicated, heterogeneous data. Random Forest's versatility in handling different feature types and their interactions makes it a great fit for this dataset, which combines structured numerical and categorical features like property size, location, and number of bedrooms with unstructured data like noise levels and possibly textual descriptions or images. One of its main benefits is that it can capture detailed, non-linear correlations between attributes, which is essential for modelling housing pricing in situations where complex interactions exist between things like noise pollution, proximity to important places, and property characteristics. Furthermore, Random Forest offers insights into feature relevance, assisting you in determining which traits (noise, for example). Even though Random Forest is a great place to start, you should compare its results with those of other algorithms, such as XGBoost or Gradient Boosting Machines (GBM). These models work well for capturing interactions between features and can provide higher accuracy in some situations. To prevent overfitting, they might need to be carefully adjusted, nevertheless. In combination with structured data, you can experiment with neural networks like as Convolutional Neural Networks (CNNs) for images or Recurrent Neural Networks (RNNs) for text if you plan to add unstructured data like text or images (e.g., property descriptions or photos). In the end, you can efficiently manage complex feature interactions by starting with Random Forest, and you can make sure you're using the optimum approach for your data by comparing it afterward with other models.

© 2024 IJSRET

5. Model Training

Divide the dataset into training and testing sets using a standard 80/20 ratio to start the model training phase. This indicates that the model will be trained on 80% of the data, with the remaining 20% set aside for evaluating the model's performance on untested data. Make sure that during this procedure, the aspects related to noise level are included in addition to other crucial details like the size, location, and number of rooms of the property. By doing this, the model will be able to understand the connections between these attributes and home values. It's crucial to use cross-validation techniques once the training set is prepared to make sure your model performs effectively when applied to new data. A popular technique is called k-fold cross-validation, in which k subsets, or folds, of the training data are divided, and the model is trained on k-1 folds while the remaining fold is utilised for validation. To ensure that every fold is utilised for validation precisely once, this operation is performed k times. By preventing the model from being unduly tailored to a single subset of data, this strategy helps prevent overfitting and offers a more accurate prediction of the model's performance on fresh, unseen data.

6. Evaluation

In the model evaluation phase, using Mean Squared Error (MSE) and R-squared (R^2) provides effective metrics to assess the performance of your housing price prediction model. MSE measures the average of the squares of the errors, quantifying the average squared difference between predicted values and actual prices, with lower MSE indicating better performance. It is calculated using the formula are the predicted prices, and n is the number of observations. In contrast, R-squared assesses the proportion of variance in the dependent variable (house prices) explained by the independent variables (features, including noise levels). R2 is calculated by the formula to predict values. R^2 values range from 0 to 1, with values closer to 1 indicating that the model explains a larger proportion of variance. By comparing the MSE and R^2 of models with and without noise level features, you can determine the impact of incorporating noise data on prediction accuracy. A lower MSE and a higher R^2 in the model with noise features would indicate their significant contribution to enhancing predictive capability, ultimately guiding further improvements to the model if necessary.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

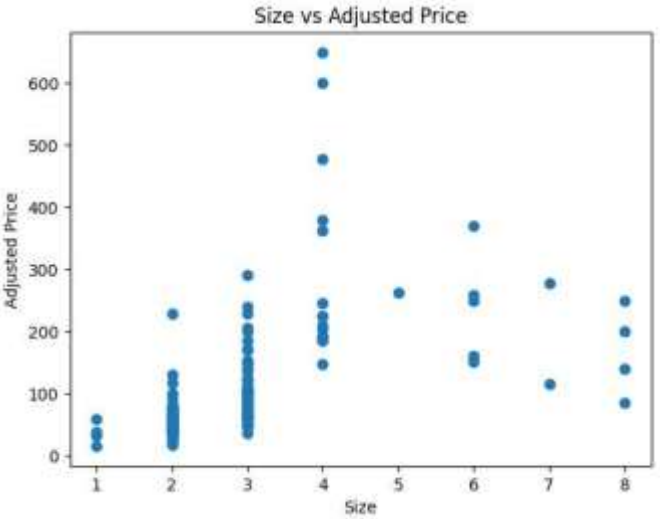
7. Visualization

During the visualisation stage, you can make graphs that effectively show the correlation between noise levels and home prices using programs like Matplotlib. To begin, create scatter plots to illustrate the relationship between noise level and property prices. To show the various noise levels on a single graph, you can either make separate plots for each noise category (such as low, medium, and high) or utilise colour coding. Additionally useful for visualising the spatial distribution of noise in various places and its relationship to house prices are heatmaps. Box plots, which compare home values across noise level bands to show how median and interquartile price ranges change in places with different noise pollution, could be another helpful visual aid. Visualising the data using this way can help understand how noise pollution affects property prices on a deeper level. These graphics can show how noise levels affect things clearly and draw attention to patterns that might not be immediately obvious from the raw data. Including noise level data with predictive model will help better understand the elements influencing Bangalore real estate values and make more accurate forecasts about house prices.

IV. CONCLUSION

By providing the total square feet and bhk, we were able to obtain the adjusted pricing in the results.It was finally the estimated cost of the house.

We used two criteria to determine the project's accuracy. R^2 and mean square error measures.



Error on Mean Squared: 6212.39012391568
R-squared is the 0.15979270145888003
R-squared error should be less than 0.5,but since we took less noise readings,we got 0.15; if we took more readings, the R-squared error would have been better accuracy.

REFERENCES

1. P. -Y. Wang, C. -T. Chen, J. -W. Su, T. -Y. Wang and S. -H. Huang, "Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism," in IEEE Access,

© 2024 IJSRET

vol. 9, pp. 55244-55259, 2021, doi: 10.1109/ACCESS.2021.3071306.

2. A.Jaswanthi, P. Jaya Sri, S. Rajendranadh, N. Snehith Prediction of House Pricing Using Machine Learning with Python IJIRT Volume 8, Issue 4 2021.
3. X. Fu, T. Jia, X. Zhang, S. Li, and Y. Zhang, "Do street level scene perceptions affect housing prices in Chinese megacities? An analysis using open access datasets and deep learning," PLoS ONE, vol. 14, no. 5, 2019.
4. MS.A.Vidhyavani1,O.Bhargav Sathwik2,Hemanth.T3, Vishnu Vardhan Yadav .M4. Housing Price Prediction Using Machine Learning IJCRT Volume 9, Issue 11 2021.
5. Z. Peng, Q. Huang, and Y. Han, "Model research on forecast of second hand house price in Chengdu based on XGboost algorithm," in Proc. IEEE11th Int. Conf. Adv. Infocomm Technol. (ICAIT), Oct. 2019.
6. S. Law, B. Paige, and C. Russell, "Take a look around: Using street viewand satellite images to estimate house prices," ACM Trans. Intell. Syst.Technol., vol. 10, no. 5, pp. 1–19, Nov. 2019.

© 2024 IJSRET