REAL ESTATE PRICE PREDICTION WEB APPLICATION

Submitted by: Group 2

BHARATH MARTURU
KAVYA MANNAVA
KOWSALYA MURALIDHARAN
SHIVANAND PULAKANTI
SHIVANI BOMMIDI

TABLE OF CONTENTS

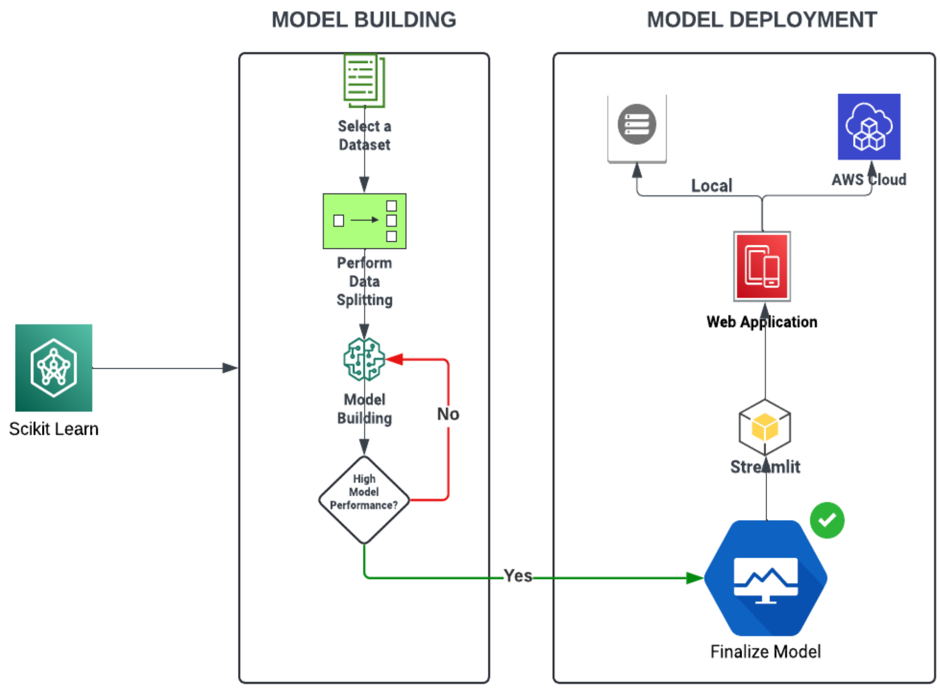
1. PROBLEM DEFINITION

The world of real estate is intricate and always changing, making it tough for buyers, sellers, and agents to accurately determine the value of properties. The motivation behind this project is to address this complexity by developing an advanced Real Estate Price Prediction System. Our goal is to create a powerful machine learning model that can deliver accurate property valuations based on data. By giving everyone involved in real estate transactions access to this valuable information, we aim to help them make better decisions when buying, selling, or investing in properties.

The proposed solution is unique because it includes a wide range of specific property features and qualities in the predictive model. Instead of relying on a limited set of basic attributes, the model will consider detailed information like the size of the property, number of bedrooms and bathrooms, floor level (for apartments), and other crucial factors. By considering these specific property details, the machine learning model aims to provide a better understanding of what determines a property's value. This could involve factors like the amount of living space for a growing family, the luxury and appeal of a high-rise view, the convenience of multiple bathrooms, or other special property features.

The Real Estate Price Prediction System aims to provide precise and dependable property value estimates through a thorough valuation method. It aims to assist buyers, sellers, and agents in making informed decisions during real estate transactions. The system's objective is to equip all involved parties with data-driven information to confidently navigate the complexities of the real estate market.

2. SYSTEM DESIGN



The Real Estate Price Prediction System consists of several key components that work together to provide accurate property valuations. The figure above illustrates the high-level overview of the system design.

2.1 Data

The study utilizes a real estate dataset obtained from the Open-Source website, Kaggle, with a focus on Delhi, India. The dataset comprises 3,555 rows of data, which provides a substantial amount of information to analyze. The dataset consists of various property attributes, which were used for model training.

To ensure accurate analysis, the collected data underwent preprocessing, including handling missing values, outliers, and feature scaling. In addition, categorical variables were transformed into binary representation using the One-hot encoding technique. Numerical variables were standardized using the StandardScaler method to ensure accurate analysis. By implementing these measures, the accuracy of the analysis is enhanced, and any potential distortions or inaccuracies that could arise are effectively mitigated.

2.2 Modeling

The system employs a Random Forest Model with 500 decision trees as base estimators selected after rigorous testing and comparison with other algorithms like Linear Regression, XGBoost, and LASSO etc., Hyperparameter tuning is performed to optimize the model's performance and enhance prediction accuracy.

2.3 Deployment

The trained model is deployed using Streamlit, a popular framework for building interactive web applications. The application is hosted on AWS, leveraging the cloud platform's scalability and reliability.

2.4 User Experience (UX)

The Streamlit interface provides an intuitive and user-friendly experience for users to input property details and receive instant price predictions. The application is designed to be accessible to a wide range of users, including potential property buyers, sellers, and real estate agents.

The choice of technologies, such as Streamlit and AWS, was based on their ability to support the system's requirements. Streamlit allows for the creation of an engaging and interactive web application, while AWS provides the necessary infrastructure for deployment and scalability.

3. MACHINE LEARNING COMPONENT

The Machine Learning model is trained on a diverse dataset that includes a wide range of property attributes, such as location within the city, number of bedrooms, bathrooms, and balconies, built-up area (in sqft), property age, furnishing type, storeroom, maid room and floor of the apartment. Prior to implementing the model on the training dataset, we engaged in preprocessing steps to ensure the data was suitably prepared. This involved transforming categorical variables into binary representation through the OneHotEncoder technique. Additionally, numerical variables underwent standardization using the StandardScaler method. These preprocessing steps were crucial in ensuring the data was appropriately formatted and ready for model training.

3.1 Initial Iteration

Various Models like Random Forest, Linear Regression, XG Boost, Support Vector Machines, etc., were initially developed to establish a baseline performance. After evaluating the Mean Absolute Error (MAE) and R-Squared scores, we identified XGBoost and Random Forest as the top-performing models due to their superior accuracy. Consequently, we chose to conduct hyperparameter tuning for these models to further optimize their performance.

MODEL	R^2 SCORE	MAE
XGBOOST	0.88	0.52
RANDOM FOREST	0.85	0.58
EXTRA TREES	0.83	0.62
GRADIENT BOOSTING	0.85	0.63
DECISION TREE	0.75	0.68
MLP	0.76	0.78
ADABOOST	0.73	0.90
SVM	0.68	0.95
RIDGE	0.73	0.96
LINEAR_REG	0.73	0.96
LASSO	0.00	1.57

Table. Model parameters and results for first iteration

3.2 Final Iteration

Following the implementation of hyperparameter tuning on both Random Forest and XGBoost models, we observed the MAE and R-squared Score for Random Forest and XG Boost as below:

MODEL	R^2 SCORE	MAE
RANDOM FOREST	0.90	0.46
XGBOOST	0.90	0.47

Table. Model parameters and results for first iteration

After analyzing the provided values, it became clear that the Random Forest model showed superior performance compared to the alternatives based on MAE value. The model was evaluated using cross-validation techniques to assess its generalization ability and robustness.

Therefore, we proceeded to update the Random Forest model with optimized hyperparameters. The best parameters obtained for the Random Forest model are as follows:

RandomForestRegressor (*n_estimators=500, *, criterion='squared_error', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=1.0, max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, ccp_alpha=0.0, max_samples=None, monotonic_cst=None*)

The trained Random Forest model on AWS (Amazon Web Services) Interface takes user inputs for various property attributes and provides accurate price predictions. The model's ability to capture complex relationships and handle many features makes it well-suited for real estate.

4. SYSTEM EVALUATION

4.1. Validation and Performance Evaluation

Our evaluation strategy was meticulous, leveraging a comprehensive testing framework to validate the reliability and robustness of our predictive models. We conducted a comparative analysis of several machine learning algorithms and determined the Random Forest model to be the most effective due to its superior capability in managing complex data relationships. Key metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were utilized to measure predictive accuracy. These metrics provided essential insights into the average errors and the variability of predictions, offering a clear assessment of model precision and effectiveness.

4.2. System Limitations

Despite the robustness of our model, we identified several limitations that could impact its applicability:

- **Data Diversity:** Our model faces limitations due to the composition of the dataset, which features properties from certain segments of the market. This lack of representation across a broader range of property types and locations could introduce biases into the predictions, affecting the model's accuracy and general applicability.
- **Feature Inclusion:** The model currently lacks variables that reflect dynamic market conditions and macroeconomic factors, potentially limiting its accuracy under rapidly changing market conditions.
- **Scalability and Efficiency:** There are significant challenges in scaling the model to process larger datasets or deliver real-time predictions, which could affect computational efficiency and response times negatively.

4.3. Presentation of Results

Results were thoughtfully visualized through intuitive charts and graphs within our Streamlit application, designed for optimal user engagement and information clarity. We ensured detailed communication of performance metrics and their practical implications. Each model's accuracy and error metrics were transparently displayed, providing stakeholders with a comprehensive understanding of the model's capabilities and limitations. This approach not only highlighted the effectiveness of our model but also pinpointed critical areas for potential enhancement.

4.4. Recommendations for Improvement

- **Data Enrichment:** Enhancing the dataset to cover a broader spectrum of property types and incorporating more geographic diversity is crucial for improving the model's applicability and accuracy.
- **Advanced Feature Integration:** Integrating real-time market data, economic indicators, and neighborhood-specific dynamics could significantly refine the model's responsiveness to fluctuations in the market.

Algorithm Optimization: Investigating more sophisticated machine learning techniques, such as ensemble methods or neural networks, could potentially offer better performance and greater predictive reliability.

4.5. Conclusion

Our system evaluation confirms that while the Real Estate Price Prediction System demonstrates robust performance within its current scope, substantial opportunities exist for its enhancement, particularly in data diversity, feature integration, and model scalability. Addressing these areas will equip the system better to meet a diverse user base's need and adapt to the dynamic real estate market. Continued refinement, rigorous testing, and user feedback are essential to ensure the system remains robust and effective in delivering accurate property valuations. Future developments should also consider incorporating feedback mechanisms to capture user satisfaction and areas for further refinement, ensuring the system evolves in line with user expectations and market demands.

5. APPLICATION DEMONSTRATION

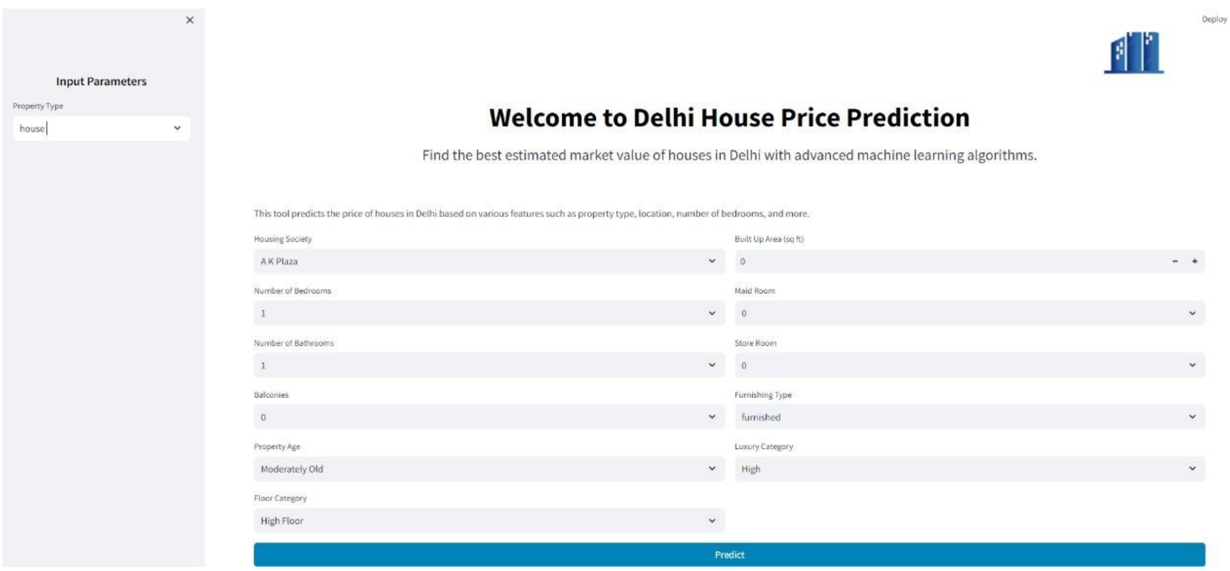
The Real Estate Price Prediction System is accessible through a user-friendly web application interface. Users can simply launch the application by accessing the provided URL, which leads them to an interactive Streamlit interface.

The interface allows users to input various property details, such as square footage, number of bedrooms and balconies, location within the city, and floor number of the apartment. Once the user has entered the necessary information, the "Predict Price" button to receive a price prediction for that particular case.

The decision to develop a web application interface using Streamlit was based on several factors. Firstly, a web application provides easy accessibility for users without the need for any additional software or setup. Secondly, Streamlit's intuitive and interactive features enhance the user experience, making it simple for users to input data and view results.

One can use the application as below:

- **Launch the Web Application:** Access the URL to launch the web application interface.
- **Input Property Details:** Enter relevant property details such as type of property, location within the city, square footage, number of bedrooms and balconies, and floor number of the apartment into the designated input fields.
- **Generate Prediction:** After entering the necessary information, click on the "Predict Price" button to generate an instant price prediction based on the provided parameters.
- **View Results:** The predicted price will be displayed on the screen, providing users with valuable insights into the estimated value of the property.
- **Experiment and Iterate:** Users can further explore the application by adjusting input parameters and generating additional predictions as needed. This iterative process allows users to gain a deeper understanding of how several factors influence property prices and refine their decision-making accordingly.



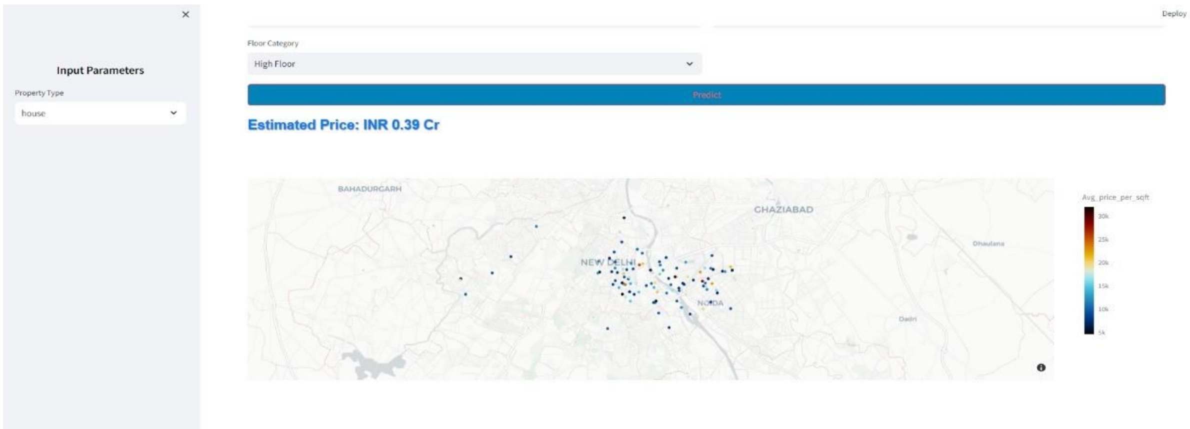


Fig. User interface of the application

[Demo Link]

6. REFLECTION

Looking back on the development of the Real Estate Price Prediction System, several aspects worked well, while others presented challenges and opportunities for improvement.

What worked well:

- The choice of the Random Forest algorithm proved to be effective in capturing the complex relationships within the real estate data and providing accurate price predictions.
- The use of Streamlit for the user interface allowed for the creation of an intuitive and interactive web application, enhancing the user experience.
- The deployment of the system on AWS ensured scalability, reliability, and fast response times for users.
- Embracing these new tools Streamlit and AWS EC2 allowed us to broaden our technical expertise and explore innovative solutions, empowering us to deliver more robust and sophisticated projects.
- The team dynamics were positive, with effective collaboration and communication among team members.

What did not work and areas for improvement:

- The limited availability of comprehensive real estate datasets posed challenges in training the model on a wider range of property attributes and locations.
- The model's ability to account for subjective factors, such as neighborhood dynamics and property conditions, was limited due to the complexity of quantifying such factors.
- The user interface could be further enhanced with additional features, such as visualizations and comparative analysis, to provide users with more insights and context.

If given unlimited time and resources, several enhancements could be made to the application:

- Expanding the dataset to include a broader range of property attributes and locations, enabling the model to provide more accurate predictions across different markets.
- Incorporating advanced techniques, such as deep learning and natural language processing, to capture subjective factors and sentiment analysis from property descriptions and reviews.
- Integrating additional data sources, such as economic indicators and demographic information, to provide a more comprehensive understanding of the real estate market.
- Enhancing the user interface with interactive visualizations, personalized recommendations, and market trends analysis to offer users a more immersive and informative experience.

Moving forward, the team plans to continue refining the model and expanding its capabilities. Potential next steps include collaborating with real estate professionals to gather additional data and insights, exploring partnerships with real estate platforms to integrate the system, and conducting user testing to gather feedback, improve the application's usability, and value proposition.

7. BROADER IMPACTS

The Real Estate Price Prediction System has the potential to revolutionize the way property valuations are conducted in the real estate industry. By providing accurate and data-driven price predictions, the system can assist buyers, sellers, and agents in making informed decisions, ultimately leading to a more transparent and efficient market.

The application, while constructed to help users make reasonable decisions regarding transactions on real property, presents room for exploitation from some of the users regarding the capability of forecasting prices. This might be used to engage in speculative buying and selling, which can lead to market volatility and market bubbles. For example, if the valuation of past properties is done through discriminatory practices, then there is a risk that the model may inadvertently perpetuate these biases, giving biased results in the real world.

To mitigate these risks, the application has been designed in a way that the prediction algorithm would have complete transparency, informing users of what the factors being taken into consideration are and, more importantly, what limits the model has. This will help users to understand the predictions' context and help them make informed decisions.

As the Real Estate Price Prediction System continues to evolve, it is essential to engage with stakeholders, including real estate professionals, regulatory bodies, and consumer advocacy groups, to address potential concerns and ensure the system is used ethically and responsibly.

8. REFERENCES

Streamlit documentation: https://docs.streamlit.io/

AWS documentation: https://docs.aws.amazon.com/

Scikit learn documentation: https://scikit-learn.org/stable/getting_started.html

Factors Affecting Real Estate Market | Macroeconomics: Youtube

Random Forest algorithm: Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.