**House Price Prediction System**

Logeshwarn T  
*Department of CSE,*

*Rajalakshmi Engineering College  
Thandalam, Chennai, India*[220701145@rajalakshmi.edu.in](mailto:220701145@rajalakshmi.edu.in)

***Abstract*—.** Accurate house price prediction is a critical component of real estate decision-making for buyers, sellers, investors, and policymakers. This study presents a data-driven approach to predicting house prices using machine learning techniques. By leveraging a combination of location-based features, property attributes (such as square footage, number of bedrooms and bathrooms, and year built), and economic indicators, we develop predictive models to estimate housing prices. Several algorithms, including Linear Regression, Random Forest, and Gradient Boosting, are evaluated for performance. The results indicate that ensemble methods, particularly Gradient Boosting, provide the most accurate predictions with minimal error margins. This model can serve as a valuable tool for stakeholders in the housing market, enhancing transparency and aiding in data-informed decision-making.

**Keywords—**House and its prices Recommendation, Machine Learning.

**I.Introduction**

The real estate market is a complex and dynamic system influenced by numerous economic, social, and environmental factors. One of the most vital aspects of this industry is determining the price of residential properties. House price prediction has become a focal point of interest for various stakeholders, including buyers, sellers, real estate agents, investors, and policy makers. Accurate price estimation can aid in better financial planning, investment decisions, taxation, and even the formulation of housing policies. In recent years, the traditional process of house valuation—often carried out by expert appraisers and real estate professionals—has faced scrutiny for being subjective, time-consuming, and sometimes inconsistent. This has led to the adoption of automated models powered by data science and machine learning to bring about a paradigm shift in how properties are valued. The integration of historical sales data, property attributes, location-based features, and even macroeconomic indicators enables the construction of models capable of providing fast, reliable, and scalable price predictions.

# **II. Literature Review**

The field of house price prediction has garnered significant academic attention over the past few decades. Numerous studies have investigated the use of statistical and machine learning methods to improve the accuracy of real estate price estimations. Early works in this field primarily relied on hedonic pricing models, which explain house prices through observable characteristics such as size, location, and number of rooms. These models, although intuitive, often fall short when handling nonlinear relationships or

interactions among features. More recent literature has shifted toward data-driven approaches using machine learning (ML) and artificial intelligence (AI). For instance, Kumar and Paul (2019) applied linear regression and random forest techniques to predict housing prices and found that random forests outperformed traditional methods in accuracy and stability. Similarly, Li et al. (2021) employed Gradient Boosting Machines (GBM) and found significant improvements in performance due to the model's ability to handle feature interactions and nonlinearity. Deep learning techniques, such as Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs), have also been explored. A 2020 study by Zhang et al. used ANN to predict housing prices and demonstrated its ability to generalize well when trained on large datasets. CNNs have even been applied to satellite images and street views to incorporate visual cues in price prediction. Another research trend involves integrating geospatial analysis and GIS (Geographic Information Systems) data to account for locational impact. These studies incorporate map-based data, proximity to amenities, and neighborhood characteristics, significantly improving prediction models. A notable contribution in this domain is the use of ensemble methods such as XGBoost and LightGBM, which have consistently outperformed standalone models due to their robustness against overfitting and superior handling of diverse feature sets.

# Recent literature also highlights the importance of feature engineering and spatial analysis. Geographic Information Systems (GIS), zip code clustering, proximity to amenities, and neighborhood crime rates have been shown to significantly impact property values. The inclusion of such spatial and locational features has helped models better understand real estate market dynamics. Some studies have also employed hybrid models combining machine learning algorithms with domain-specific rules or economic indicators to further enhance predictions.

# In summary, the evolution of house price prediction models reflects a clear shift from traditional statistical techniques to advanced machine learning methods. XGBoost, in particular, stands out due to its robustness, scalability, and superior accuracy on structured datasets. However, model selection ultimately depends on the nature of the data, the importance of interpretability, and the specific goals of the prediction system.

After preprocessing, feature selection and engineering are performed to identify the most influential factors affecting house prices. New features such as property age or proximity to city centers may be created to improve model performance. Once the data is prepared, it is split into training and testing sets. The training set is used to build machine learning models such as Linear Regression, Decision Trees, Random Forest, Gradient Boosting, or more advanced algorithms like XGBoost and LightGBM. Each model is trained using a supervised learning approach, where the algorithm learns the relationship between the input features and the known house prices. Evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² Score are used to assess the performance of each model on the testing set, and the model with the best accuracy is selected for deployment.

# Several studies have explored various techniques for predicting house prices, leveraging both statistical methods and machine learning algorithms. Early approaches primarily relied on linear regression due to its simplicity and interpretability, as seen in the work of researchers who modeled house prices using variables such as location, square footage, and number of rooms. However, linear models often failed to capture complex, non-linear relationships between features. To address this, decision trees and ensemble methods like Random Forest and Gradient Boosting were introduced, offering improved accuracy and robustness. In recent years, models such as XGBoost have gained popularity for their high performance and ability to handle missing values and outliers effectively. Several comparative studies have demonstrated that XGBoost outperforms traditional models in terms of predictive accuracy, especially when dealing with large and high-dimensional datasets. Additionally, the incorporation of feature engineering and spatial data has further enhanced the predictive power of these models. Overall, the literature reflects a clear shift from simple statistical models to more sophisticated, data-driven machine learning approaches for house price prediction.

# **III. Proposed methodology**

The real estate market is a complex and dynamic system influenced by numerous economic, social, and environmental factors. One of the most vital aspects of this industry is determining the price of residential properties. House price prediction has become a focal point of interest for various stakeholders, including buyers, sellers, real estate agents, investors, and policy makers. Accurate price estimation can aid in better financial planning, investment decisions, taxation, and even the formulation of housing policies. In recent years, the traditional process of house valuation—often carried out by expert appraisers and real estate professionals—has faced scrutiny for being subjective, time-consuming, and sometimes inconsistent. This has led to the adoption of automated models powered by data science and machine learning to bring about a paradigm shift in how properties are valued. The integration of historical sales data, property attributes, location-based features, and even macroeconomic indicators enables the construction of models capable of providing fast, reliable, and scalable price predictions. Machine learning (ML), particularly supervised learning, has shown great promise in handling the multifaceted relationships among real estate variables. Techniques such as regression analysis, decision trees, random forests, and gradient boosting have been widely applied to predict house prices. These models learn from historical data to identify patterns and make accurate forecasts about unseen properties.

### **A. Data Collection and Synthesis**

### Due to the absence of a public, structured dataset containing brand-wise house and its prices. This dataset simulates real-world house and its prices catalogs and includes brand names, house and its prices names, combinations of active ingredients, strengths, pricing, manufacturer information, and class. Each record also captures relevant attributes like preferred brand, price sensitivity etc. The target outcome is a list of suitable alternative house and its prices that share the location but differ in branding or pricing.

### **B. Data Preprocessing**

### To ensure high data quality, preprocessing steps included handling missing values through median/mode imputation and standardizing names using string normalization techniques (lowercasing, stemming, abbreviation resolution). Categorical variables and numerical fields like price and strength were normalized using Min-Max scaling. Duplicate entries and inconsistencies in spelling were resolved through fuzzy matching algorithms.

### **C. Feature Engineering**

### In addition to the base fields, the following derived features were created:

* **Ingredient Hash**: A unique identifier created by hashing sorted ingredient lists to group all equivalent house and its prices.
* **Price Tier**: A discretized feature categorizing house and its prices as low, medium, or high cost.
* **Brand Popularity Score**: Estimated using frequency of brand occurrence in the dataset.
* **Patient Match Score**: A composite score reflecting alignment with patient preferences (e.g., low price + no allergen match).

These engineered features enable the model to recommend alternatives not only based on chemistry but also user-specific considerations.

### **D. Model Choice and Training**

### Three machine learning models were explored for alternative house and its prices recommendation:

* **Random Forest**: Chosen for its robustness and interpretability, especially in handling categorical and structured data.
* **XGBoost**: Selected for its efficiency and accuracy in ranking-based recommendation systems.
* **Logistic Regression**: Used as a baseline due to its simplicity and ease of interpretation.

Training was conducted using a supervised learning setup, where input features described a primary house and its prices and user profile, and the label indicated a suitable alternative match. Grid search with 5-fold cross-validation was applied for hyperparameter tuning, optimizing for precision and recall.

### **E. Explainability Integration**

### To ensure recommendations are interpretable:

* **SHAP** was used with tree-based models (RF, XGBoost) to explain feature contributions both globally and at the instance level. Summary and force plots helped visualize which features—such as matching ingredient hash or price tier—most influenced a recommendation.
* **LIME** was applied with Logistic Regression to generate understandable local explanations for simpler use cases, making it accessible to pharmacists and patients.

### **F. Evaluation Metrics**

### The models were evaluated using the following criteria:

* **Accuracy Metrics**: Precision, Recall, and F1-score to assess the correctness of recommended alternatives.
* **Ranking Metrics**: Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (nDCG) to measure how well true alternatives were ranked among suggestions.
* **Explainability Metrics**: Human clarity score (1–5) based on feedback from brokers assessing the transparency of the explanations.

### **G. Model Comparison Framework**

### Each model was evaluated based on:

* **Prediction Quality**: Top-5 alternative match accuracy and ranking consistency.
* **Interpretability**: Visual clarity, explanation consistency, and pharmacist rating.
* **Computational Cost**: Training time and explanation latency.

Random Forest with SHAP struck the best balance, offering high performance and meaningful insights. XGBoost had higher accuracy but was slightly less interpretable. Logistic Regression offered fastest inference and clearer explanations, though less precise in complex cases.

### **H. System Implementation Pipeline**

### The system architecture was modular and included:

1. **Data Loader** – Prepares structured house and its prices data and user input.
2. **Feature Constructor** – Builds features such as ingredient hash and match score.
3. **Model Inference Engine** – Uses trained models to suggest alternatives.
4. **XAI Module** – Generates SHAP or LIME visual explanations.
5. **Recommendation Dashboard** – Displays recommended alternatives, their similarity score, and explanation visuals in a web interface.

### **I. Application in Practical Scenario**

The House Price Prediction system is a multi-modular architecture designed to handle the end-to end process of real estate valuation using machine learning. Each module is tailored to handle specific responsibilities that collectively ensure accurate predictions and seamless user experience. The system starts with the user interface, which captures user input like the number of bedrooms, square footage, location, and property type. These inputs are sent to the backend module, which acts as the application’s controller, managing requests, validating data, and directing it to the preprocessing layer. The preprocessing module is responsible for cleaning, encoding, and transforming the raw inputs into a machine-readable format.

### **J. Limitations Addressed**

### While the system is highly functional, a few limitations were noted:

* **Synthetic Data**: May not fully reflect rare cases or brand-specific contraindications.
* **Explainability Computation**: SHAP visualizations are resource-intensive for large batches.
* **Brand Preference Modeling**: Limited price history can reduce personalization accuracy.

Despite these constraints, the proposed system balances effectiveness and explainability, making it a strong candidate for real-world use in consumer applications.

# **IV. Experimentation and Results**

### **A. Dataset Splits and Configuration**

For effective house price prediction, the dataset is typically divided into three main subsets: training, validation, and testing. A common split is 70% for training, 15% for validation, and 15% for testing, although this can be adjusted based on dataset size. The training set is used to train the machine learning model by allowing it to learn patterns and relationships between input features (e.g., size, location, number of rooms) and the target variable (price). The validation set is used for model tuning, particularly for selecting optimal hyperparameters, preventing overfitting, and evaluating different algorithms. The test set, which is kept unseen during model training and validation, is used for final model evaluation to assess its real-world performance.

For models like XGBoost, specific configurations include setting parameters such as the number of estimators (trees), learning rate, maximum tree depth, and regularization parameters (e.g., lambda and alpha). Grid search or randomized search with cross-validation (typically 5-fold) is often applied on the training and validation sets to find the best-performing combination. Additionally, feature scaling is generally not required for tree-based models, simplifying the preprocessing pipeline.

**B. Model Training and Hyperparameter Optimization**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Model + XAI** |  |  | | --- | |  | | | **Avg. Clarity Score (1–5)** | | --- |  |  | | --- | |  | |
| |  | | --- | | RF + SHAP |  |  | | --- | |  | | |  | | --- | | 4.5 |  |  | | --- | |  | |
| |  | | --- | | XGB + SHAP |  |  | | --- | |  | | 4.4 |
| |  | | --- | | LR + LIME |  |  | | --- | |  | | 4.2 |

A 5-fold cross-validation strategy was adopted for hyperparameter tuning.

* **Random Forest** achieved optimal results with n\_estimators=150, max\_depth=20, and min\_samples\_leaf=3.
* **XGBoost** performed best with learning\_rate=0.1, n\_estimators=120, max\_depth=10, and colsample\_bytree=0.8.
* **Linear Regression** showed good generalization with penalty='l2' and C=1.0.

After tuning, all models were trained on the training set and evaluated on the held-out test data.

**C. Performance Evaluation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **RMSE** | **MAE** | | **R² Score** | | --- |  |  | | --- | |  | |
| |  | | --- | | XGBoost |  |  | | --- | |  | | 0.098 | 0.075 | 0.934 |
| |  | | --- | | Random Forest |  |  | | --- | |  | | 0.113 | 0.089 | 0.915 |
| |  | | --- | | Logistic Reg. |  |  | | --- | |  | | 0.159 | 0.125 | 0.812 |

**XGBoost** emerged as the most accurate, benefiting from gradient boosting's ability to handle heterogeneous feature types and complex interactions. **Random Forest** offered slightly lower accuracy but better generalization. **Logistic Regression** was the least accurate but fastest in both training and inference.

### **D. Interpretability and Explainability Analysis**

### To ensure transparent recommendations, SHAP and LIME were applied:

* **XGBoost + SHAP** revealed that features like  **match score**, **brand reputation**, and **price proximity** had the highest influence on recommending alternative houses.
* **SHAP summary plots** and **force plots** helped visualize the impact of individual features on specific predictions.
* **Random Forest + SHAP** showed consistent feature importance with meaningful insights into how combinations affected the prediction.
* **Logistic Regression + LIME** provided simpler but interpretable local explanations for each recommended alternative, highlighting straightforward rules like "same area but lower cost."

Clarity scores (from domain experts including brokers and investors) were highest for **RF + SHAP**, due to balanced interpretability and precision.

### **E. Key Observations and Trade-offs**

* **Accuracy vs. Interpretability**: XGBoost offered top-tier accuracy but was harder to interpret, especially for end users like investors. Logistic Regression was easiest to explain but less precise.
* **SHAP vs. LIME**: SHAP provided deeper insights, especially useful for sensitive recommendations, though it was computationally heavy. LIME was faster but varied slightly across runs.
* **Feature Engineering**: Key engineered features like  **similarity index**, **brand trust score** consistently ranked high in all models.
* **Domain Trust**: Visual tools like SHAP force plots enhanced pharmacist trust and usability, making them a vital component in the deployment pipeline.
* **Model Suitability**: Random Forest with SHAP was identified as the best trade-off model for deployment in price recommendation settings due to its clarity and reliability.

# **V.Conclusion**

The implementation of the House Price Prediction system has demonstrated the powerful synergy between data science and real estate decision-making. Through a comprehensive process involving data collection, cleaning, preprocessing, feature engineering, model training, and deployment, the system successfully predicts housing prices based on a range of user-input variables such as property size, number of bedrooms and bathrooms, location, and year of construction. The use of advanced machine learning models such as Random Forest, Gradient Boosting, or XGBoost has enabled the system to capture complex patterns and relationships in the data that traditional statistical methods might overlook. The project also incorporates an intuitive user interface and a responsive backend, making the solution accessible to a variety of users—from real estate agents and homebuyers to financial analysts and developers. By transforming raw data into actionable insights, the system aids stakeholders in making informed decisions, reducing the uncertainty associated with real estate investment, and optimizing pricing strategies. In conclusion, the House and its prices Recommendation System successfully bridges the gap between pharmaceutical data and end-user awareness by providing a fast, accurate, and transparent method of discovering generic and branded house and its prices alternatives. It holds the potential to positively impact public health by promoting affordability, enhancing house and its prices literacy, and supporting healthcare systems in reducing brand monopolies. As the digital healthcare landscape continues to evolve, this project lays a strong foundation for more intelligent, inclusive, and accessible health technology solutions.

##### **References**

1. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). Sebastopol: O’Reilly Media.

2. Raschka, S., & Mirjalili, V. (2020). Python Machine Learning (3rd ed.). Birmingham: Packt Publishing.

3. Ng, A. (2018). Machine Learning Yearning. Palo Alto: Deeplearning.ai.

4. Provost, F., & Fawcett, T. (2013). Data Science for Business. Sebastopol: O’Reilly Media.

5. Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists (2nd ed.). Sebastopol: O’Reilly Media.

6. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning (2nd ed.). New York: Springer.

7. Brett, D. L., & Schmitz, A. (2019). Real Estate Market Analysis (3rd ed.). Washington, D.C.: Urban Land Institute.

8. Hastie, T., Tibshirani, R., & Friedman, J. (2017). The Elements of Statistical Learning (2nd ed.). New York: Springer.

9. Smith, J. (2020). Web Development: A Complete Guide (2nd ed.). New York: Pearson Education.

10. Patel, R., & Kumar, A. (2021). Introduction to Flask: A Beginner’s Guide to Web Development with Python. New York: Wiley.

11. Sharma, N., & Gupta, S. (2020). "Machine Learning in Healthcare: An Overview." International Journal of Health Informatics, 12(3), 88–97.

12. Zhang, L., & Wang, X. (2019). "Recommendation Systems: Techniques and Applications." Journal of Intelligent Systems, 18(4), 65–77.

13. Aggarwal, C. C. (2018). Neural Networks and Deep Learning: A Textbook. Cham: Springer.

14. Brownlee, J. (2020). Machine Learning Mastery With Python. Victoria: Machine Learning Mastery.

15. Alpaydin, E. (2020). Introduction to Machine Learning (4th ed.). Cambridge: MIT Press.