Project Report: Advanced Housing Price Prediction with Feature Engineering and SHAP Analysis

Rana Mandal (Student No: 2041577)

1. Introduction

Housing price prediction plays a vital role in the real estate market by helping buyers, sellers, and investors make informed decisions. This project employs advanced machine learning techniques, feature engineering, and model interpretability tools (SHAP and LIME) to predict housing prices accurately. We utilize a Random Forest Regressor model, augmented by SHAP (SHapley Additive exPlanations), to achieve high predictive performance and enhance model interpretability.

2. Data Overview and Preprocessing

The dataset (Housing_Price_Dataset) includes real estate features such as SquareFeet, Bedrooms, Bathrooms, YearBuilt, Neighborhood, and Price. The target variable is Price.

- **Handling Missing Values**: Numeric columns were filled with their median values, while categorical columns were filled with their mode to ensure completeness without introducing significant bias.
- **Feature Engineering**:
    - **HouseAge** was derived as 2025 - YearBuilt, representing the age of the house.
    - **SquareFeet_per_Bedroom** and **Bedroom_to_Bathroom_Ratio** were created to capture additional relationships and improve predictive capabilities.
    - **Log Transformation** was applied to SquareFeet and Price to address skewness in their distributions.
- **Categorical Encoding**: One-hot encoding was applied to the Neighborhood column to convert categorical values into numeric features.
- **Data Normalization**: The features were scaled using StandardScaler to ensure that all features contribute equally to the model, avoiding bias due to different scales.
-

3. Model Selection and Training

Four regression models were selected for evaluation:

- **Linear Regression**: A simple and interpretable model.
- **Decision Tree Regressor**: A non-linear model suitable for capturing more complex relationships.
- **Random Forest Regressor**: An ensemble method that aggregates multiple decision trees to improve accuracy and reduce overfitting.
- **Gradient Boosting Regressor**: Another ensemble model that builds trees sequentially to improve model performance.

The dataset was split into an 80% training set and a 20% testing set. For the **Random Forest Regressor**, hyperparameter tuning was done using GridSearchCV to optimize key parameters:

- n_estimators: Number of trees (set to 100).
- max_depth: Maximum depth of trees (optimized to 10 and 20).
- min_samples_split: Minimum samples to split a node (set to 2).

The optimal configuration was found to be max_depth=10 and n_estimators=100.

4. Model Evaluation

The models were evaluated using the following metrics:

- **Mean Absolute Error (MAE)**: Average magnitude of errors in predictions.
- **Root Mean Squared Error (RMSE)**: Penalizes larger errors more than MAE.
- **R^2 (Coefficient of Determination)**: Measures the proportion of variance explained by the model.

Model	Train R^2	Test R^2	Test RMSE	Test MAE
Linear Regression	0.65	0.61	48,000	38,000
Decision Tree Regressor	0.90	0.58	50,000	40,000
Random Forest Regressor	0.92	0.67	45,000	37,000
Gradient Boosting Regressor	0.88	0.64	47,000	37,500

The **Random Forest Regressor** emerged as the best-performing model with the highest **Test R^2 (0.67)**, lowest **RMSE (45,000)**, and **MAE (37,000)**, making it the ideal choice for this task.

5. Model Interpretability

5.1 SHAP Analysis:

- **Global Interpretability**:
    - **Summary Plot**: Displays the contribution of each feature to the model's predictions, showing their influence across all test samples.
    - **Key Features**: The most influential features included HouseAge, SquareFeet_per_Bedroom, and Bedroom_to_Bathroom_Ratio, with HouseAge being the most impactful feature for predicting housing prices.
- **Dependence Plot**: Showed the relationship between HouseAge and predicted price. Older houses generally had lower prices, reinforcing the intuitive assumption about property depreciation over time.

5.2 LIME Analysis:

- **Local Interpretability**:
    - LIME was used to generate explanations for individual predictions. It provided insights into the top features influencing specific predictions.
- **Visualization**: LIME was visualized for selected test instances, highlighting the impact of different feature values on individual predictions, enhancing transparency and understanding of model behavior.

6. Conclusion and Future Work

6.1 Key Findings:

- **Feature Engineering**: Creating new features like SquareFeet_per_Bedroom and Bedroom_to_Bathroom_Ratio significantly improved model accuracy.
- **Random Forest Regressor**: This model outperformed others in terms of predictive accuracy, demonstrating its effectiveness for capturing complex patterns in the data.

- **SHAP and LIME**: These tools provided valuable insights into model decision-making, enhancing transparency and helping to identify the most impactful features.

6.2 Future Work:

- **Further Feature Engineering**: Investigate additional feature interactions and transformations to improve model performance.
- **Model Experimentation**: Test other models such as XGBoost or more advanced Gradient Boosting methods to evaluate their performance against Random Forest.
- **Hyperparameter Optimization**: Fine-tune additional Random Forest parameters to boost performance.
- **Ensemble Methods**: Explore stacking models for improved predictive performance.
- **Deep Learning**: Explore deep learning techniques like **TabNet** for potential improvements in prediction accuracy.

6.3 Tools Used:

- **SHAP**: Applied for global model interpretability, highlighting the influence of each feature on the predictions.
- **LIME**: Used for local interpretability to explain individual predictions.
- **Random Forest**: The primary model used for regression due to its robustness and accuracy.
- **PyTorch Tabular**: Explored as a deep learning approach for tabular data to further improve predictive performance.