HOUSE PRICE PREDICTION MODEL

A Project Report submitted for Cognitive Computing (UCS420)

by

Abhishek Panchal 102317167

Kaushik Gupta 102317265

Submitted to

Mr. Sukhpal Singh

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology

Patiala, Punjab, India

March 2025

Introduction and Problem Statement

The real estate market is highly dynamic, with house prices influenced by factors like location, economic conditions, and property characteristics. Accurately predicting house prices is vital for homebuyers, real estate agents, investors, and policymakers. However, traditional pricing methods often fail due to their reliance on simplistic assumptions, leading to inaccurate estimates.

This project aims to address this real-world challenge by developing a house price prediction model that overcomes the limitations of traditional pricing methods. Unlike conventional approaches that rely on simple assumptions, our model captures complex relationships between various housing attributes, making it more robust and accurate.

The dataset used for this project is sourced from Kaggle's **House Prices – Advanced Regression Techniques,** which provides detailed information on housing characteristics and neighbourhood conditions. This data forms the foundation for building a predictive system that can enhance decision-making in real estate valuation.

Dataset Link - https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

Data Exploration and Preprocessing

The dataset used for this project originally consists of 1,460 samples and 79 columns. The refined dataset used for training the machine learning models consists of **1,387 records and 15 selected columns**.

The final 15 columns are: **OverallQual, GrLivArea, GarageCars, BsmtUnfSF, TotalSF, Age, RemodAge, LotFrontage, Qual_LivArea** (Overall Quality × Living Area), **Garage_Qual** (Garage Cars × Overall Quality)**, TotalSF_Sq, Age_Sq, Neighborhood_MedianPrice** (median house price by neighborhood), **Neighborhood_Cluster** and **SalePrice** (Target variable).

Preprocessing steps:

1. **Feature Engineering:** Created features like *TotalSF, Age, RemodAge, Neighborhood_MedianPrice*, and clustered neighborhoods into 5 groups based on *Neighborhood_MedianPrice.*
2. **Outlier Removal:** Outliers in critical columns like *GrLivArea* and *TotalSF* were removed using Local Outlier Factor (LOF), refining the dataset to 1,387 records.
3. **Advanced Features:** Created interaction (*Qual_LivArea, Garage_Qual*) and polynomial features (*TotalSF_Sq, Age_Sq*) to capture non-linear complex relationships.
4. **Missing Value Handling:** Imputed missing values in *LotFrontage* with the median.
5. **Scaling:** Numerical features were standardized using *StandardScaler* to improve model convergence.

Model Implementation & Evaluation

To improve prediction accuracy and capture complex relationships in housing data, the following machine learning models were implemented and evaluated:

1. **Random Forest Regressor:** Robust against overfitting and capable of handling complex, nonlinear relationships. It is ideal for leveraging key property and neighborhood features that impact house prices.
2. **Decision Tree Regressor:** Useful for understanding feature importance and baseline comparisons. Although simpler, it provides interpretability and captures key decision paths influencing house prices.
3. **XGBoost Regressor:** Known for high accuracy and efficiency, XGBoost is well-suited for handling large datasets, reducing prediction errors, and capturing complex feature interactions.

The training process involved the following key steps:

1. **Data Splitting:** The dataset (1,387 records and 15 features) was split into **training (75%)** and **test (25%)** sets.
2. **Hyperparameter Tuning:** RandomizedSearchCV was used to optimize critical hyperparameters to improve accuracy and generalization.
3. **Training and Model Selection:** Each model was trained on the training set, and the best model was selected based on its performance on the test set.

The models were evaluated using 3 metrics: **Root Mean Squared Error (RMSE)**, **R^2 Score** and **Mean Absolute Error (MAE).**

Results and Insights

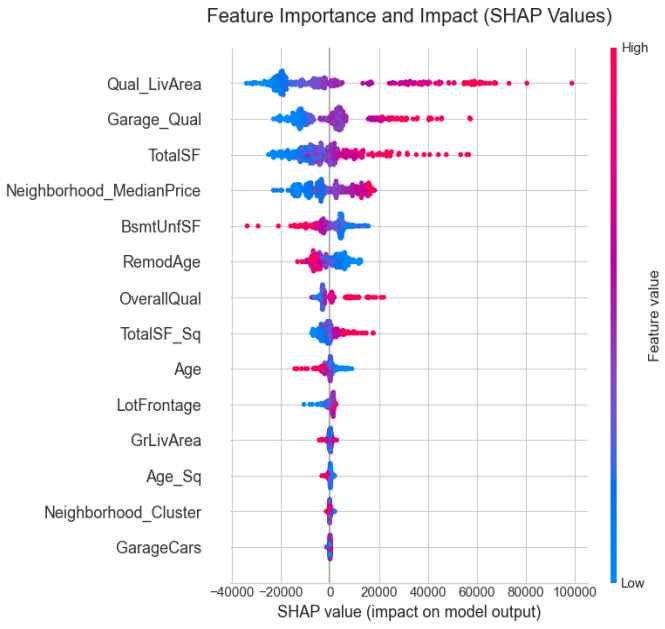
Performance Results:

- **Random Forest:** RMSE = $25989.97, R^2: 0.8820, MAE: $17245.62
- **Decision Tree:** RMSE = $30016.95, R^2: 0.8427, MAE: $20309.45
- **XGBoost:** RMSE = $24608.19, R^2: 0.8943, MAE: $16305.76

Based on these results, **XGBoost** emerged as the best-performing model. Additionally, SHAP analysis was used to interpret the model's predictions and assess the influence of key features.

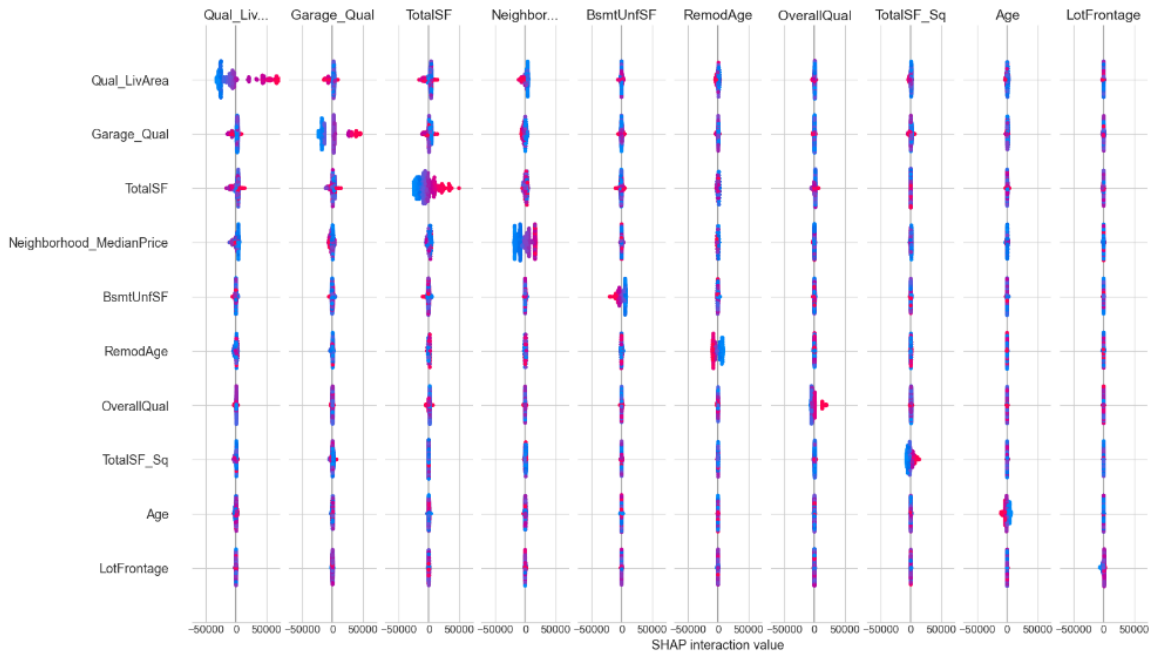
Key insights from the SHAP plots include:

- **Model Accuracy:** Most predicted prices closely follow the diagonal line, indicating strong model performance, with XGBoost performing best, especially for higher-priced homes.
- **Strong Mid-Range Accuracy:** Models predict mid-range prices ($100,000–$300,000) accurately, suggesting a stable and competitive market in this segment.
- **Underprediction for High-End Properties:** Higher-priced homes (above $400,000) show underprediction, indicating that luxury pricing complexities may not be fully captured by the models.

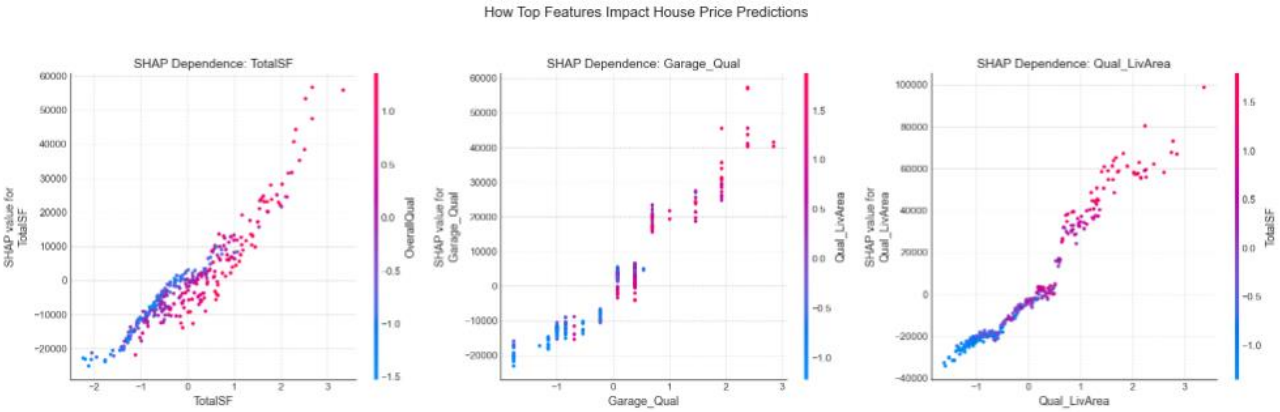


- **Top Price Drivers:** *Qual_LivArea, Garage_Qual*, and *TotalSF* have the most significant impact on house prices, with higher values leading to higher predicted prices.
- **Neighborhood Effect:** *Neighborhood_MedianPrice* shows that local market trends strongly influence pricing, reflecting geographic price segmentation.
- **Renovations Enhance Property Value:** The positive impact of *RemodAge* highlights the importance of modernizing homes to improve their market value.

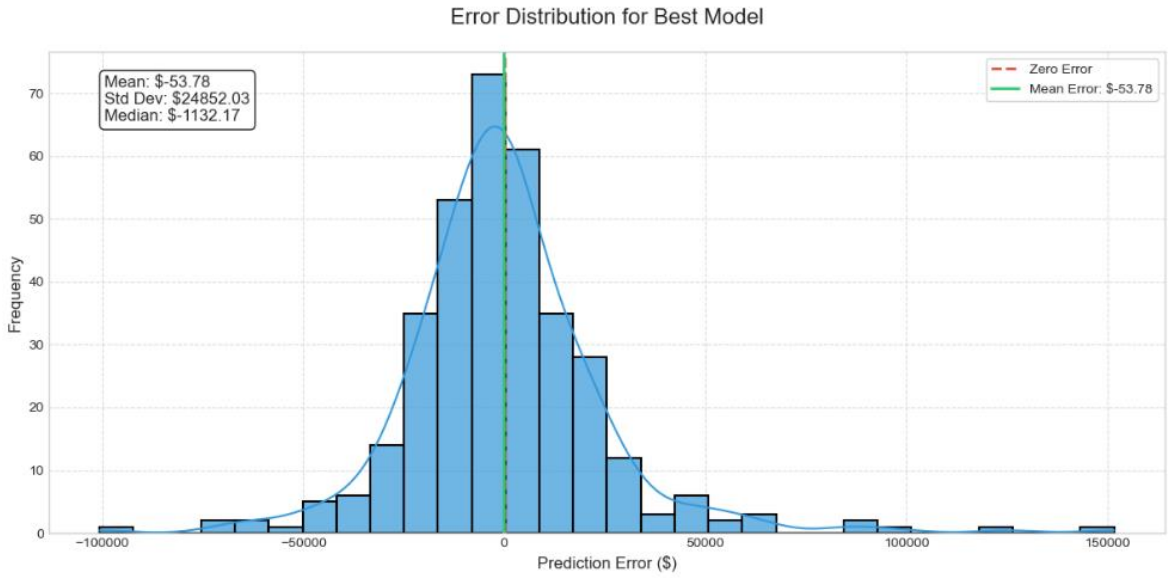
- **Garage Consideration:** Garage quality has more influence than garage size, emphasizing quality over quantity in buyer preferences.



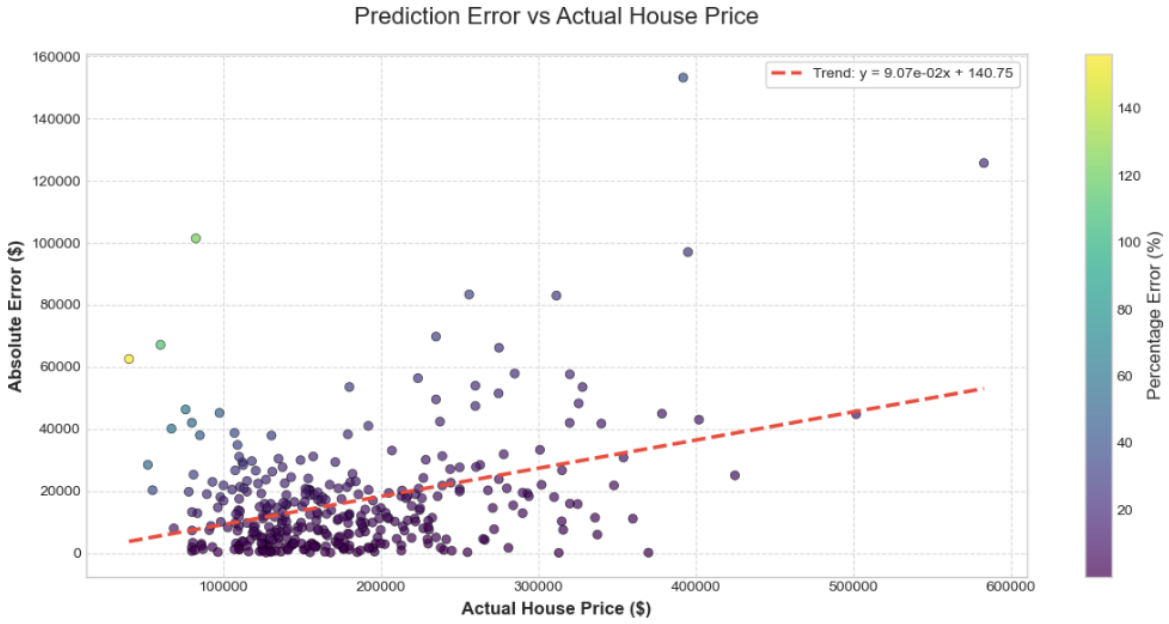
- **Dominant Features:** *TotalSF, OverallQual*, and *Neighborhood_MedianPrice* strongly impact house prices, reflecting the importance of size, quality, and location.
- **Renovation and Age Dynamics:** *RemodAge* and *OverallQual* interaction indicates that well-renovated older homes retain higher value.
- **Preference for Finished Spaces:** *BsmtUnfSF* interacts negatively with *TotalSF*, reflecting a preference for finished living areas over unfinished basements.



- **TotalSF Impact:** Larger homes significantly raise house prices, with a sharp increase in value at higher square footage. Interaction with *Qual_LivArea* amplifies this effect, reflecting the premium placed on both size and living area quality.
- **Garage_Qual Impact:** Higher garage quality boosts predicted prices, with noticeable jumps at higher ratings. Interaction with *Qual_LivArea* further increases the value for well-built properties.
- **Qual_LivArea Impact:** Improved living area quality has a nonlinear, dramatic effect on prices, especially at top quality levels. Combined with *TotalSF*, it strongly drives property value.



- The distribution of errors closely resembles a **normal distribution**, indicating that the model's predictions are generally **unbiased** and **well-calibrated**. This is often a sign of a **well-performing regression model**.
- **Mean error** ($10.18) and **Median error** (-$1,120.81) are close to zero, indicating **balanced** and **accurate** predictions.



- **Positive Correlation:** Absolute error tends to increase with higher actual house prices, as indicated by the upward trend line.
- **Small Errors Dominate:** Most errors cluster below $20,000, especially for homes priced under $200,000, suggesting solid accuracy in the lower price range. While high priced properties show more variability.

Challenges & Future Improvements

The following challenges were identified -

- **Feature Interaction Complexity:** SHAP interaction plots indicate complex relationships between key features like *TotalSF, Qual_LivArea*, and *Garage_Qual*. While these derived features improved accuracy, the model may still not fully capture nuanced interactions between categorical variables like neighborhoods and overall quality.
- **Outliers & Rare Properties:** High prediction errors for luxury homes due to limited data representation and complex, atypical features.
- **High Error Variability:** Increased prediction variance for higher-priced properties, indicating uneven model performance across price ranges, potentially due to insufficient training data for that price range.

Further Improvements –

- **Data Augmentation:** Generate synthetic data to balance underrepresented price segments (luxury homes).
- **Stacked or Ensemble Models:** Implement LightGBM, CatBoost or model stacking (combining Random Forest, XGBoost, and linear regression) to balance bias and variance.
- **Robust Validation:** Perform Bayesian or Genetic Algorithm-based hyperparameter tuning and employ K-fold cross-validation with stratification to ensure robustness across different price ranges.