Advanced House Price Prediction Using Machine Learning and Ensemble Methods

Anirudh N.
Department of MCA, Chanakya University, Bengaluru, Karnataka 562110, India
Corresponding author: anirudhn.mca24@chanakyauniversity.edu.in

Impana P.
Department of MCA, Chanakya University, Bengaluru, Karnataka 562110, India
Corresponding author: impanap.mca24@chanakyauniversity.edu.in

ABSTRACT

Accurate prediction of house price is vital for real estate stakeholders, enabling data-driven decision-making in the complex market. This study further extends the already published work of Yang [1] by applying machine learning models like Linear Regression, Random Forest, and Decision Tree – to the Kaggle House Prices dataset and introducing advanced ensemble techniques: Gradient Boosting model (XGBoost) and Stacked Model. Upon using optimized parameters, the Stacked Model achieved the lowest Mean Squared Error then all the other competitors taken in consideration. This indicates the prediction of House Price can be best when used with more than one type of model thus giving a more intricate solution

1. INTRODUCTION

House price forecasting is a pillar of real estate research, informing investors, policymakers, and homeowners. The intricacy of the market, underpinned by economic, geographic, and social dynamics, necessitates advanced modelling to identify nonlinear trends [2]. Machine learning has been found effective in this application, using large data sets to enhance predictive accuracy [3].

This work extends Yang's [1] work, where he used Linear Regression, Random Forest, and Decision Tree models to forecast house prices. We reproduce these models and augment the methodology with Gradient Boosting (XGBoost) and a Stacked Model, seeking to minimize prediction errors and offer an all-round performance comparison. Our results demonstrate the promise of ensemble models in enhancing real estate analytics.

2. DATA AND MODEL

2.1 Data Collection

The data set, sourced from the Kaggle "House Prices: Advanced Regression Techniques" competition [10], includes 2919 property's records with 79 features, including LotArea (land area), OverallQual (construction quality), GrLivArea (grand living area), and GarageCars (garage cars capacity). The target variable SalePrice, ranges from $34,900 to $755,000 with a mean of $180,921. The data set separated into training set (1460 records) and test sets (1459 records), is verified to be high quality [1].

2.2 Data Preprocessing

According to Yang [1], we performed extensive preprocessing for ensuring the quality of data. The missing values were filled with medians for numeric features and modes for categorical features. The outliers in vital features (e.g., LotFrontage, GrLivArea) were eliminated based on the Interquartile Range (IQR) approach. Numeric features were standardized using StandardScaler, while categorical variables were one-hot encoded, leading to over 200 features. Both the train set, and test set were then merged for uniform preprocessing and then divided to maintain the integrity of evaluation.

Feature selection privileged features with strong correlation to SalePrice, such as OverallQual and GrLivArea, found by Yang [1]. Additional feature engineering was considered but not investigated to focus on computational efficiency.

2.3 Models

We implemented five models: three from Yang [1] (Linear Regression, Random Forest, Decision Tree) and two advanced ensemble methods (XGBoost, Stacked Model).

- **Linear Regression**: A baseline model with Ridge regularization, tuned for the alpha parameter via cross-validation.
- **Random Forest**: An ensemble of decision trees, tuned for n_estimators and max_features to enhance generalization.
- **Decision Tree**: A single tree model, with max_depth tuned to balance complexity and overfitting.
- **XGBoost**: A gradient boosting model that corrects errors iteratively, tuned for n_estimators, max_depth, and learning_rate.
- **Stacked Model**: Combines predictions from Linear Regression, Random Forest, Decision Tree, and XGBoost using a Linear Regression meta-learner.

The target variable (SalePrice) was log-transformed to address skewness, with predictions exponentiated for evaluation, consistent with [1].

3. ANALYSIS OF RESULT

3.1 Data Visualization

Data visualization is a key mechanism in demystifying the Kaggle House Prices dataset, taking raw data, and representing it in forms that are understandable and reveal underlying structures, distributions, and relationships. Through the application of a sequence of visualization techniques, we investigated the nature of the dataset, checked for preprocessing steps, and evaluated model performance, leveraging the work of Yang [1] for initial visualizations. The below figures provide an overview of the data and results of the model, using packages like Matplotlib and Seaborn to make it clear and precise.

The below figures provide an overview of the data and results of the model, using packages like Matplotlib and Seaborn to make it clear and precise.

Target variable distribution, SalePrice, is central to the knowledge of the dataset structure. From Figure 1, a histogram using a kernel density estimate (KDE) shows a right-skewed distribution with sale prices ranging from $34,900 to $755,000 and a mean value of $180,921. The skewness, as is usual in real estate markets, justifies the log-transformation employed in modelling to stabilize variance and improve model fit.

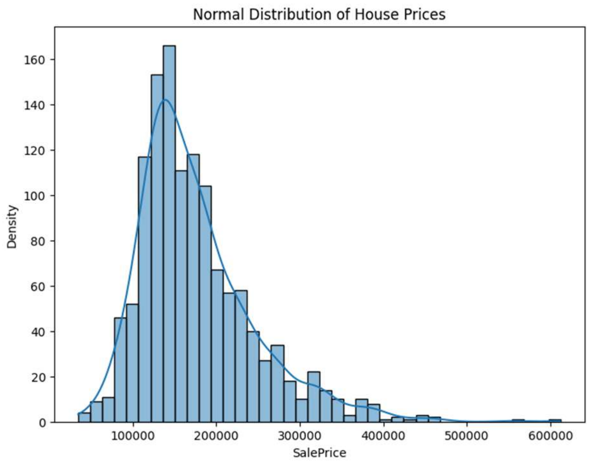

Figure 1: *Histogram of SalePrice with a kernel density estimate, showing a right-skewed distribution. The x-axis represents SalePrice ($), and the y-axis represents density.*

Correlation analysis confirmed key features influencing SalePrice, including OverallQual, GrLivArea, and GarageCars, aligning with Yang [1].

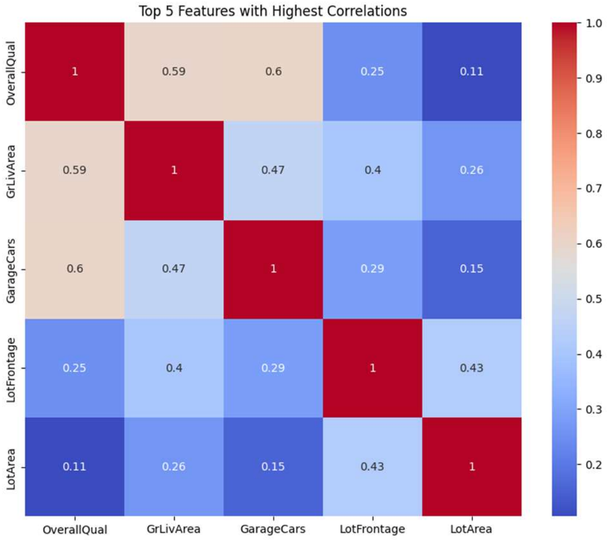

Figure 2: *Heatmap of correlation coefficients for the top 5 features, with values from -1 to 1. Strong positive correlations (e.g., OverallQual, GrLivArea) are displayed in warm colours.*

3.2 Forecasting

Linear Regression

Linear Regression: Ridge Regression was applied with the best alpha of 15.264 and had a cross-validation error of 0.0125 and an MSE of 315,488,573. It considers linear relationships, with L2 regularization preventing overfitting in high-dimensional space. Prominent features such as OverallQual had strong coefficients, with interpretability guaranteed. Its performance, remarkably close to the Stacked Model, points to its strength for linear patterns but constrains the ability to catch nonlinear interactions [1, 4].

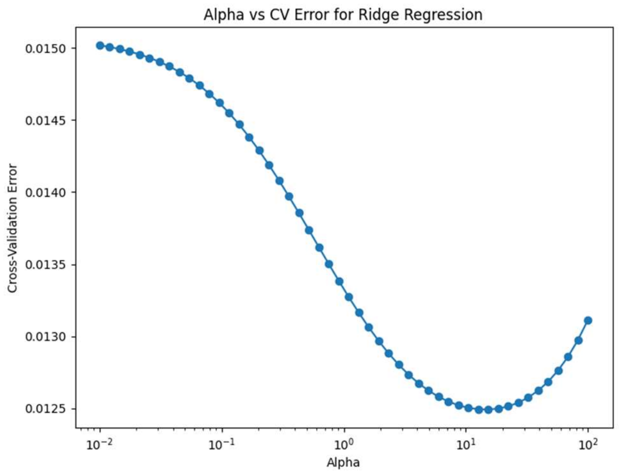

Figure 3: *Plot of alpha (x-axis, logarithmic scale) versus cross-validation error (y-axis). The error minimizes at alpha ≈ 15.264.*

The model achieved an MSE of 315,488,573, demonstrating effectiveness for linear relationships but limitations with complex patterns, as noted by Yang [1].

Random Forest

Random Forest: With 300 estimators and max_features set at 0.3, it achieved a cross-validation error of 0.0163 and MSE of 494,597,064. This ensemble approach mitigates overfitting through bagging, performing well on nonlinear data but performing worse than Linear Regression [1, 3].

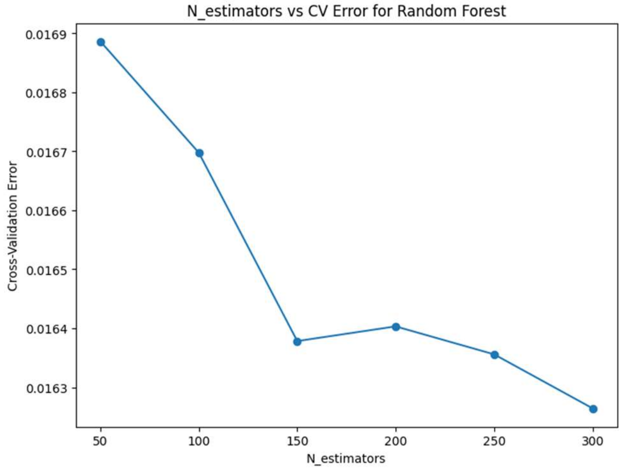

Figure 4: *Plot of n estimators (x-axis) versus cross-validation error (y-axis). The error is lowest at n_estimators = 300.*

The MSE of 494,597,064 reflects Random Forest's ability to handle nonlinearity, though it underperformed compared to other models.

Decision Tree

Decision Tree: With the best max_depth of 6, it had a cross-validation error of 0.0334 and a high MSE of 1,066,305,569. Being simple helps interpretability but causes overfitting, so it is the worst-performing model [1, 6].

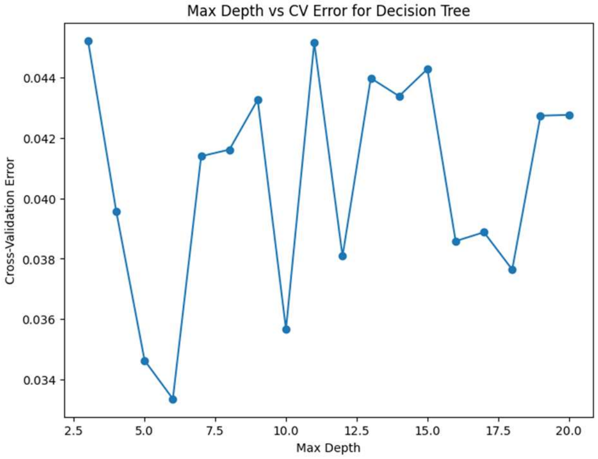

Figure 5: *Plot of max depth (x-axis) versus cross-validation error (y-axis). The error minimizes at max_depth = 6, increasing thereafter due to overfitting.*

The MSE of 1,066,305,569 indicates interpretability but significant overfitting, consistent with Yang's observations [1].

XGBoost

XGBoost: Optimized with n_estimators=150, max_depth=3, and learning_rate=0.1, it had an MSE of 377,355,872. Gradient boosting iteratively corrects mistakes, performing better than Random Forest but Linear Regression, because of proper management of intricate patterns [7].

Stacked Model

Stacked Model: Blending all the base models with a Linear Regression meta-learner, it registered the least MSE of 311,583,949. Stacking utilizes varied strengths, and hence it is most precise, best suited for modelling intricate interactions in house price data [7].

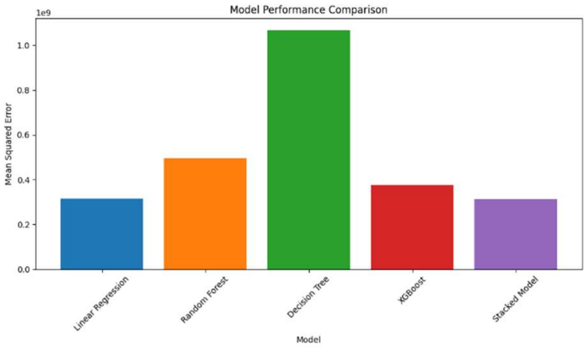

Figure 6: Bar chart comparing MSEs: Linear Regression (315,488,573), Random Forest (494,597,064), Decision Tree (1,066,305,569), XGBoost (377,355,872), and Stacked Model (311,583,949). The x-axis lists models, and the y-axis shows MSE.

4. DISCUSSION

4.1 Model Performance Comparison

Stacked Model's MSE of 311,583,949 is the lowest among all, which proves the effectiveness of ensemble learning by a multitude of diverse learners [7]. Linear Regression (MSE = 315,488,573) is comprehensible but hampered by nonlinearity [4]. XGBoost (MSE = 377,355,872) performs better than Random Forest (MSE = 494,597,064) because of iterative error refinement, justifying Yang's preference for ensemble solutions [1]. Remarkably high MSE of Decision Tree (1,066,305,569) suggests the probability of overfitting [6].

Each model is employed for specific tasks: Linear Regression for simplicity, Random Forest and XGBoost for complex patterns, Decision Trees for explainability, and Stacked Models for highest accuracy.

4.2 Factors Affecting Accuracy

Preprocessing like outlier removal and normalization mattered, Yang [1] emphasized. Feature selection for high-impact variables like OverallQual [7] enhanced performance. Adjusting parameters—alpha for Linear Regression, n_estimators for Random Forest, max_depth for Decision Tree, and some parameters for XGBoost—enhanced performance. External drivers, such as local amenities (e.g., schools, transportation) and economic cycles, influence prices [8].

Volatility in real estate markets, particularly in markets like China [9], suggests employing dynamic factors in future models [11].

5. CONCLUSION

This study validates and enhances Yang's [1] work, in which Stacked Model (MSE = 311,583,949) is the best, closely followed by Linear Regression (MSE = 315,488,573). XGBoost (MSE = 377,355,872) outperforms Random Forest (MSE = 494,597,064), and Decision Tree (MSE = 1,066,305,569) lags due to overfitting. Ensemble methods, particularly stacking, offer great improvements in prediction accuracy.

Future studies could delve into the field of deep learning, utilize real-time economic indicators, or develop adaptive models for market shifts. Such advances will refine the accuracy of house price projections, empowering real estate players.

REFERENCES

1. X. Yang, Research on House Price Prediction based on Machine Learning. *ITM Web of Conferences*, 70, 02018 (2025). https://doi.org/10.1051/itmconf/20257002018
2. B.W. Luo, Z.Y. Hong, J.Y. Wang, Application of multiple linear regression statistical modeling in house price prediction. *Computer Age*, (6), 51-54 (2020).
3. Y.L. Gong, Y.H. Yang, Construction of automatic real estate appraisal model of random forest and its comparative study. *China Asset Appraisal*, (1), 32-41 (2022).
4. S.A. Septianingrum, et al., Performance Analysis of Multiple Linear Regression and Random Forest for an Estimate of the Price of a House. *Proceedings of iSemantic*, 415-418 (2022).
5. Lei Gan, Research on second-hand house valuation model in Chongqing based on random forest model. *Ph.D. thesis, Chongqing University of Technology* (2020).
6. G.Z. Fan, S.E. Ong, H.C. Koh, Determinants of House Price: A Decision Tree Approach. *Urban Studies*, 43(12), 2301-2315 (2006).
7. L. Zhang, M. Xie, Exploration of the relationship between internal attributes of houses and house prices: based on random forest method. *Modern Business*, (22), 59-61 (2019).
8. Z. Zhang, D. Zheng, Mining analysis of objective factors affecting regional house price. *Computer Application and Software*, 36(11), 32-38, 85 (2019).
9. P.-F. Pai, W.-C. Wang, Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices. *Appl. Sci.*, 10, 5832 (2020).
10. Train, House Rates - Advanced Regression Techniques. *Kaggle* (2024, August 15), https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview.
11. Test, House Rates - Advanced Regression Techniques. *Kaggle* (2024, August 15), https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview