Machine Learning Models for Real Estate Pricing: Evaluation and Recommendation

Data Preparation Steps

1. Load the Austin housing dataset.
2. Create additional predictors for pool and renovation by searching the property description text for keywords related to pools and renovations.
3. Create an age variable by subtracting the build year from the sale year. This step also removed some redundant date-related variables.
4. Drop unnecessary variables such as detailed addresses, sale dates, and other redundant columns.
5. Convert relevant columns to factors for categorical processing.
6. Create a preprocessing recipe — remove zero-variance predictors and then dummy-encode categorical variables.
7. Prep and bake the recipe to apply transformations to the dataset.
8. Split the data into training (80%) and testing (20%) sets, stratified by latestPrice to maintain price distribution.

Stepwise Regression

1. Fitted a full linear regression model using all predictors.
2. Applied forward and backward stepwise selection using stats::step() based on AIC.
3. Selected the best subset of variables.
4. Predicted on the test set using the selected model.
5. Evaluated performance using RMSE and MAE.

Random Forest

1. Fitted a Random Forest model using randomForest with mtry ≈ $\sqrt{24} \approx 5$.
2. Removed zipcode and latestPrice to resolve categorical type issues.
3. Predicted house prices on the test set.
4. Generated and review variable importance plot.
5. Evaluated performance using RMSE and MAE.

Gradient Boosting Machine (GBM)

1. Fitted a GBM model with parameters: n.trees = 5000, depth = 4, shrinkage = 0.01, cv.folds = 5.
2. Used cross-validation to determine the optimal number of trees.
3. Predicted prices on the test set using tuned model.
4. Compared predicted vs actual prices.
5. Evaluated model performance using RMSE and MAE.

Model	RMSE	MAE
Stepwise Regression	162.34	102.07
GBM	146.22	82.12
Random Forest	138.08	77.71

Out of all the models tested, Random Forest delivered the best performance with the lowest RMSE and MAE. Therefore,we are selecting Random Forest as the final model for predicting house prices.

To support and validate the robustness of our Random Forest model, we conducted additional evaluation techniques. The key findings are:

- R-squared on Test Set: The model explained 78.6% of the variability in house prices, indicating strong predictive power for real-world data.
- Residual Analysis: Residuals were centered around zero with no strong non-linear pattern, suggesting the model was not biased and captured the structure in the data well.
- Cross-Validation (5-Fold): Consistent performance across folds confirmed the model's stability and generalizability.
- Train RMSE vs Test RMSE: The train RMSE($123.90) was slightly lower than the test RMSE($138.00), indicating a good balance between fitting the training data and generalizing to unseen data.

Conclusion: Random Forest balances accuracy and generalization well, making it the most reliable model in this study.