House Price Prediction

March 2025

Harisson Eyinda
Bylhah Mugotitsa
Molefe Maleka
Joseph Osewe

Slide Content

- Feature Selection
- Model Building
- Hyperparameter Tuning
- Business Insights and Recommendations

Problem Statement

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. A house value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all aspects that give a house its value. For example, you want to sell a house and you don't know the price which you may expect. it can't be too low or too high. To find house price you usually try to find similar properties in your neighborhood and based on gathered data you will try to assess your house price.

Understanding the data

Data

Columns: 23

Rows:21613

Source:xlsx file

Type: Secondary

Categorical types

Count: 1

Coast, condition, Yr_built, total_area. (This features were encoded for modeling purposes)

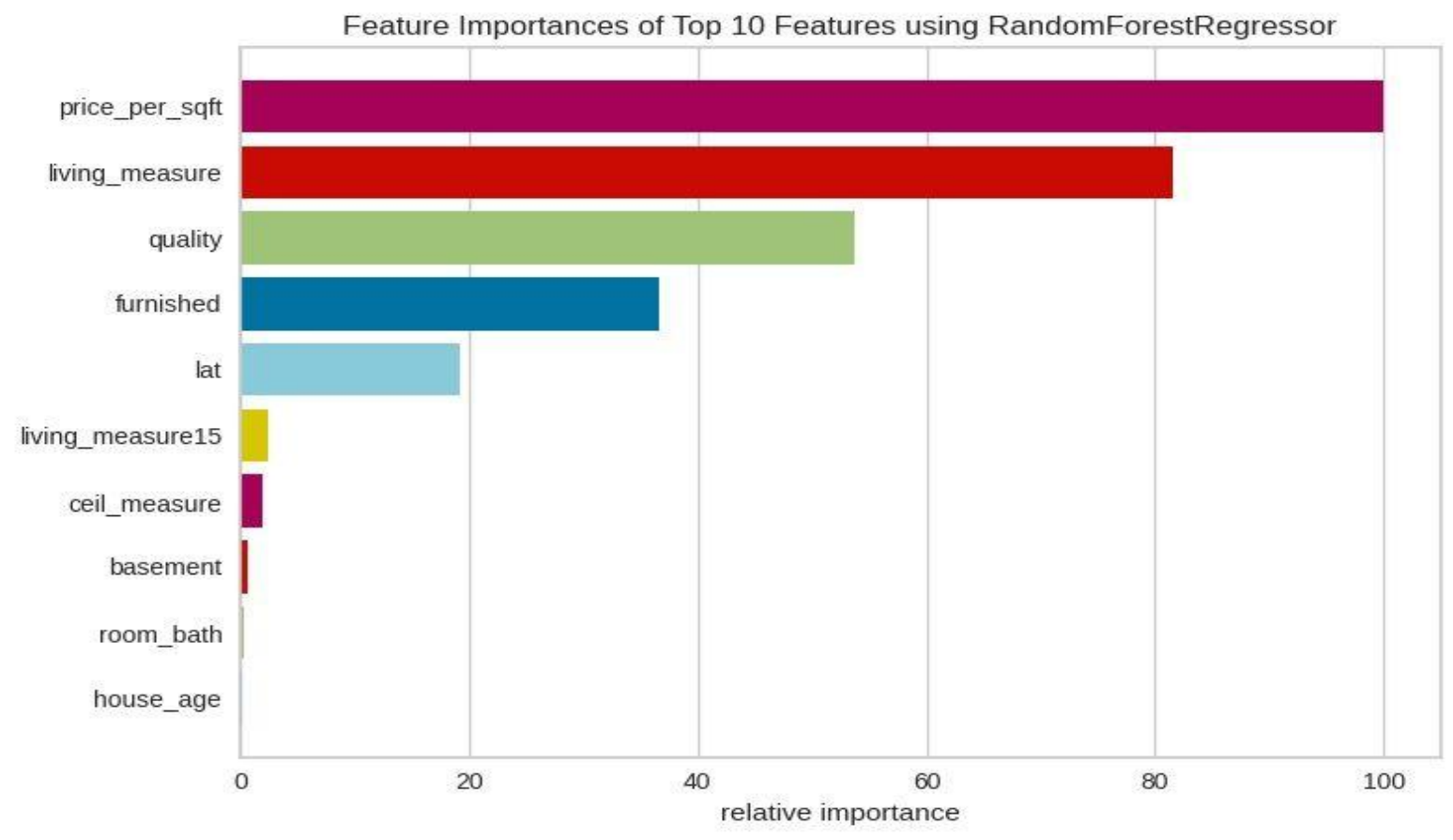
Numerical types

Count: 21

Price, room_bed, Room_bath, Living_measure, Lot_measure, Ceil, sight, Quality, ceil_measure, Basement, yr_built, Yr_renovated, Lat, long, Living_measure15, Lot_measure15, Total_area, house_age, price_per_sqft Living_lot_ratio, price, yr_built, yr_renovated

Feature Selection using Random Forest

Features with strong predictive power:



Model Building

Target Algorithms :

The following algorithms:

- Decision Tree
- Random forest
- Gradient Boost Regressor
- Xgboost Regressor

Since the algorithms are not distance based, it is not necessary to perform feature scaling

1.Decision Tree

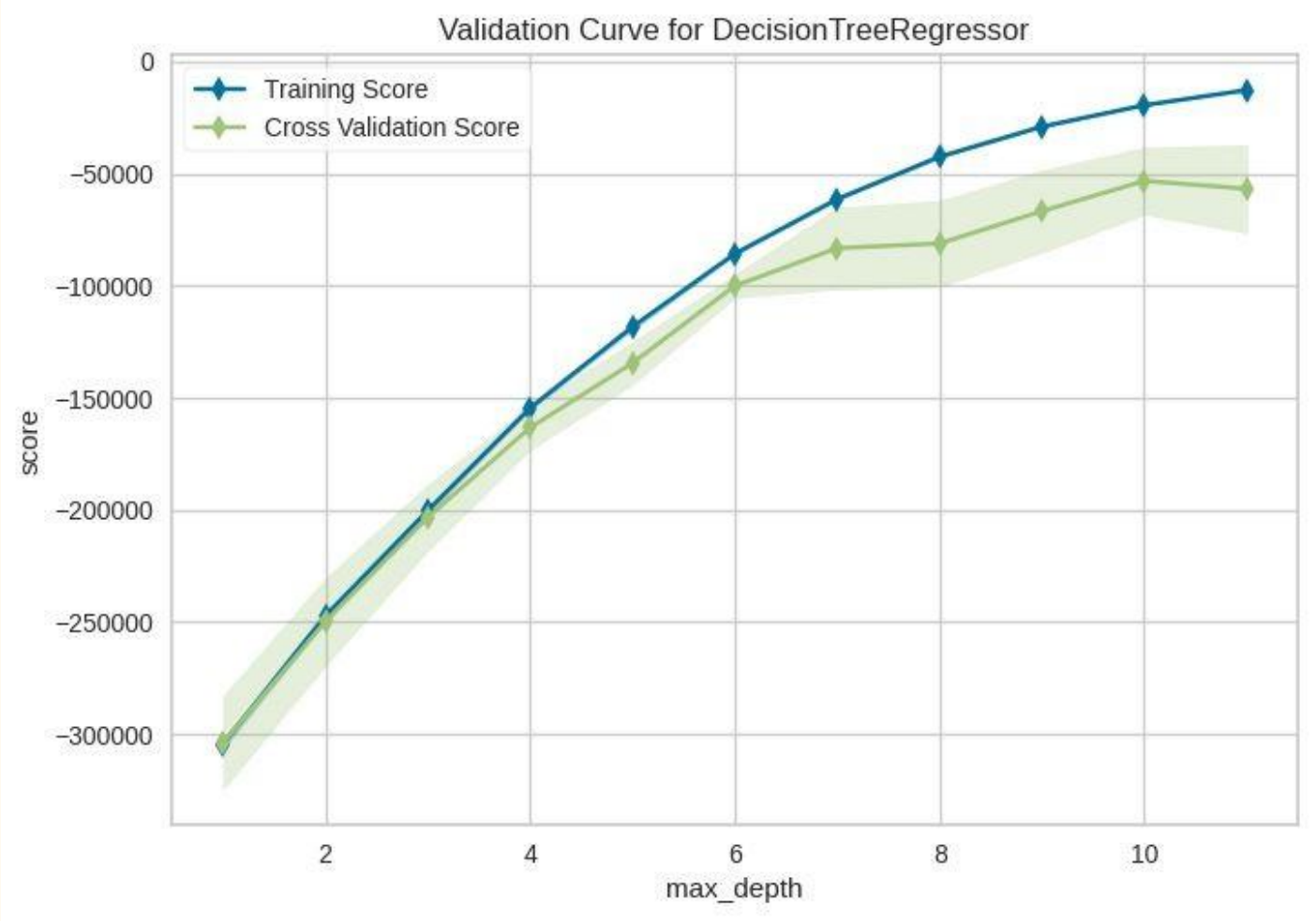
Model: Decision Tree Regressor

	MAE	RMSE	R^2
Train	0.0000	0.0000	1.0000
Test	11857.13	36715.63	0.9890

Insights: The model seems to have overfitted to the training data, as evidenced by the 0.0000 MAE, RMSE, and 1.0000 R-squared on the training set, paired with the large error values on the test set.

Attempt to Optimize

Parameter: max_depth

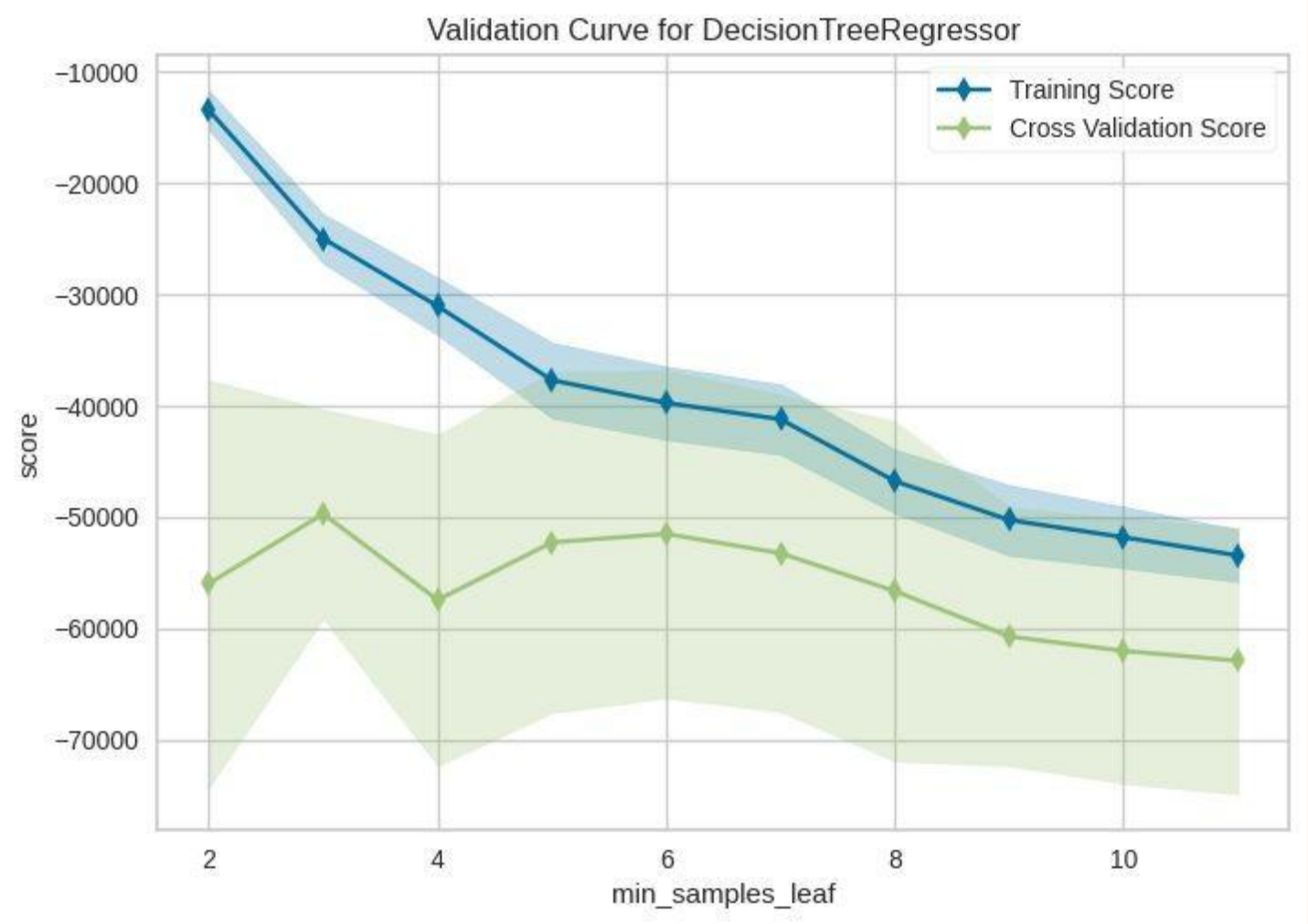


Insights:

- The training score consistently increases as max_depth increases. This is expected as deeper trees can fit the training data more closely.
- The cross-validation score also increases initially, but it starts to plateau and even slightly decrease after a certain point.
- Based on this the optimal max depth is 7

Validation Curve for

Parameter: min_samples_leaf



Insights:

- The training score consistently decreases as min_samples_leaf increases. This is expected, as higher min_samples_leaf values lead to simpler, less overfit models.
- The cross-validation score shows a similar trend but flattens out around min_samples_leaf values of 4 to 6. After this point, the cross-validation score doesn't improve significantly and might even slightly worsen.
- Therefore optimal **min_samples_leaf**: 5

After Optimization

Decision Tree Regressor

	MAE	RMSE	R^2
Train	44402.70	70151.27	0.963
Test	47970.87	79240.11	0.948

Insights: A significant improvement is observed

2. Random Forest

Model: Random Forest Model

	MAE	RMSE	R^2
Train	2116.90	11825.19	0.9990
Test	4638.72	17267.66	0.997

Insights: There is an indication of overfitting, since the model seems to be performing well on the train set compared to the test set(using the MAE and RMSE scores)

3. Gradient Boosting

Model: Gradient Boosting Model

	MAE	RMSE	R^2
Train	13928.42	20820.14	0.996
Test	14852.11	24150.65	0.995

Insights: Presence of some overfitting since the errors (MAE and RMSE) are lower in the training data compared to the tests set.The Gradient Boosting model performs well with high R-squared values on both training and test sets

4.Xgboost

Model: Xgboost Model

	MAE	RMSE	R^2
Train	6368.47	8843.60	0.9994
Test	10719.38	29952.21	0.9927

Insights: Presence of overfitting, since the model seems to perform better on the train set compared to the test set.We proceed to perform hyperparameter tuning on the three models. The decision tree was already tuned but still performed poorly compared to the other models. Hence we proceed with just the three models.

Hyperparameter Tuning

Performing hyperparameter tuning

Random Forest

Best Hyperparameters

'n_estimators': 50
'min_samples_split': 5
'min_samples_leaf': 2
'max_depth': 20
'bootstrap': True

Best Score

RMSE:
-1259884436.4394717

Performing hyperparameter tuning

Gradient Boosting

Best Hyperparameters

'subsample': 0.9

'n_estimators': 200

'max_depth': 3

'learning_rate': 0.2

Best Score

RMSE:
-711392505.1687683

Performing hyperparameter tuning

XGBoost

Best Hyperparameters

'subsample': 0.7

'n_estimators': 200

'max_depth': 5

'learning_rate': 0.05

'colsample_bytree': 1.0

Best Score

RMSE: -1018358355.2

Retraining the models using the optimal parameters

Model: Random Forest Model

	MAE	RMSE	R^2
Train	2930.37	18137.95	0.997
Test	5257.01	19590.34	0.996

Model: Gradient Boosting

	MAE	RMSE	R^2
Train	10674.457	15109.41	0.9983
Test	12382.980	22129.65	0.9959

Model: **XGBoost**

	MAE	RMSE	R^2
Train	7358.885	12060.837	0.9989
Test	8419.153	19535.153	0.9968

Insights:

- Random Forest has the lowest MAE and RMSE values for both training and testing, indicating better prediction accuracy.
- XGBoost has the highest R-squared on the training set, but the test R-squared is almost comparable to that of Random Forest.
- Gradient Boosting lags behind both Random Forest and XGBoost in terms of MAE and RMSE, while its R-squared is slightly lower than that of XGBoost and Random Forest on the test set.
- Overall, Random Forest provides the best trade-off between low error metrics (MAE, RMSE) and high R-squared on both the training and test sets, making it the most consistent and reliable model among the three.

Business Insights & Recommendations

Business Insights:

- Price per Square Foot: The inclusion of price_per_sqft as a feature suggests that the cost of a house relative to its size is a crucial factor in determining house prices. This is likely an essential variable in setting competitive pricing strategies.
- Living Space and House Quality: living_measure, living_measure15, and house_age indicate the importance of the size of the house and the age of the property. A newer, more spacious home tends to fetch a higher price, which is a common trend in the real estate market.
- Quality and Furnishing: The quality and furnished features suggest that better-quality and well-furnished homes are more valuable. This aligns with customer preferences for higher-end finishes and more luxurious living spaces.
- Geographic Location: The lat feature shows that location (likely represented by latitude) is an influential factor in house pricing, with properties in more desirable neighborhoods likely commanding higher prices.
- Additional Features (Basement, Ceiling Measure, Room Bath): Features like basement, ceil_measure, and room_bath imply that specific characteristics of the house (such as having a basement or a certain ceiling measurement) may add value to the property. These can be leveraged to highlight unique selling points for particular homes.

Recommendations:

Target Market Segmentation:

- Focus on Spacious, High-Quality Homes: Properties with larger living areas (living_measure, living_measure15) and better quality (quality) should be marketed to higher-income individuals or families who prioritize comfort and space.
- Market Furnished Homes to Premium Buyers: Homes that are furnished may appeal to individuals seeking convenience, such as those looking for rental properties or moving-in-ready homes.

Pricing Strategy:

- Utilize Price per Square Foot: Use price_per_sqft to assess market trends and price homes more competitively. This can help balance the cost of the home with market expectations based on location and size.
- Dynamic Pricing Based on Location: Since location (represented by lat) plays a significant role, tailor pricing strategies based on geographic demand. Homes in highly sought-after neighborhoods could be priced higher.

Modeling and Prediction Use:

- Consistency in Predictions: Since Random Forest provided the most reliable and consistent results, use this model to predict future house prices accurately. This can be used for pricing strategies and forecasting market trends.
- Incorporate Regular Model Updates: Given that market dynamics can change over time, it's important to periodically retrain and update models to adjust to new data, ensuring pricing models remain accurate.