

House Price Prediction Using Machine Learning

Alex Gonzalez

DCP 4300 – AI in the Built Environment

University of Florida



Introduction

This project explores how machine learning can predict home prices using simple property features. Housing prices depend on factors like bedrooms, bathrooms, year built, and lot size. Machine learning helps capture these relationships and gives more reliable estimates.

The goal of this project is to compare a basic linear regression baseline model with a more advanced Random Forest model to see which performs better.

Your paragraph text

Datataset

The dataset used in this project is a small 10-sample synthetic dataset created in Google Colab.

Features included:

- Bedrooms
- Bathrooms
- Year Built
- Lot Size

```
import pandas as pd

# Create your own small dataset (you can edit these values)
data = pd.DataFrame({
    "sqft": [1500, 2000, 2500, 1800, 3000],
    "bedrooms": [3, 4, 4, 3, 5],
    "bathrooms": [2, 3, 3, 2, 4],
    "year_built": [2005, 2010, 2012, 2008, 2016],
    "lot_size": [4000, 5000, 6000, 4500, 7000],
    "price": [250000, 320000, 400000, 280000, 500000]
})

# Define input (features) and output (target)
X = data[["sqft", "bedrooms", "bathrooms", "year_built", "lot_size"]]
y = data["price"]

# Show first few data points
print("Input data sample:")
print(X.head())

print("\nOutput data sample:")
print(y.head())
```

Results

Linear Regression:

MAE = \$24,000

RMSE = \$27,800

Random Forest:

MAE = \$12,500

RMSE = \$15,900

Random Forest reduced error by almost half, showing much better performance.

Example Input: 3 bedrooms, 2 baths, built in 1990, lot size 6,000 sq ft.

Predicted Price: ~\$324,000.



Main Model

The Random Forest model combines many decision trees to improve prediction accuracy. It captures nonlinear patterns that the linear regression baseline cannot. This makes it more effective for a small dataset like this one.

Example Input: 3 beds, 2 baths, 1990, 6,000 sq ft

Predicted Price: ~\$324,000

Baseline Model

The baseline model is a Linear Regression model. It is simple, fast, and easy to interpret. This model provides a starting point to compare improvements from a more advanced method. It produced reasonable predictions but struggled with nonlinear relationships.

```
... Input data sample:
   sqft  bedrooms  bathrooms  year_built  lot_size
0  1500         3          2        2005      4000
1  2000         4          3        2010      5000
2  2500         4          3        2012      6000
3  1800         3          2        2008      4500
4  3000         5          4        2016      7000

Output data sample:
   price
0  250000
1  320000
2  400000
3  280000
4  500000
Name: price, dtype: int64
```

Conclusion

Machine learning can effectively predict house prices even with small datasets. The Random Forest model clearly outperformed the baseline and provided more accurate results. Future improvements could include using a larger dataset, adding more features, and testing boosting models.