Predicting Housing Price Per Sqft Using Tabular Machine Learning Models

Ethan Eubanks
DCP4300 - AI in the Built Environment
University of Florida

1) Introduction and Problem Statement

Understanding housing price per square foot (price/sqft) helps buyers, sellers, and planners compare properties across different sizes and neighborhoods. Machine learning offers a fast, data-driven way to estimate this value using standard listing attributes.

Goal: Predict price per square foot using simple tabular features available at listing time.

Motivation: Provide an interpretable and efficient prediction tool for the built environment and real estate contexts.

Core Question: Which home attributes contribute most to price per square foot, and how well can simple models estimate it?

2) Dataset & Features

Dataset

- Source: King Country, WA housing sales dataset
- Total records after cleaning: 21,000+ homes
- Tabular data with numeric and categorical attributes
- Target variable formula: price_per_sqft / sqft_living

Key Features Used

- Bedrooms, bathrooms, square footage (living, basement, above grade), floors, condition, grade, lot size, year built, year renovated, Zip code

3) Methodology

Data Preparation

- Loaded a large housing dataset, removed non-predictive factors (IDs, coordinates, duplicates). Created the target variable: price_per_sqft = price / sqft_living. Created a 70/30 split with a fixed random seed for reproducibility

Feature Engineering

- Normal numeric features: beds, baths, floors, year_built, grade, condition, etc.
- One-hot encoded zip code to make location a categorical variable

Baseline Models

- Global Mean Predictor: always predicts the training mean $/sqft
- Median-by-Zip: assigns each home the median $/sqft of its zipcode
- Linear Regression: Fits a straight-line relationship between features and $/sqft

Modeling

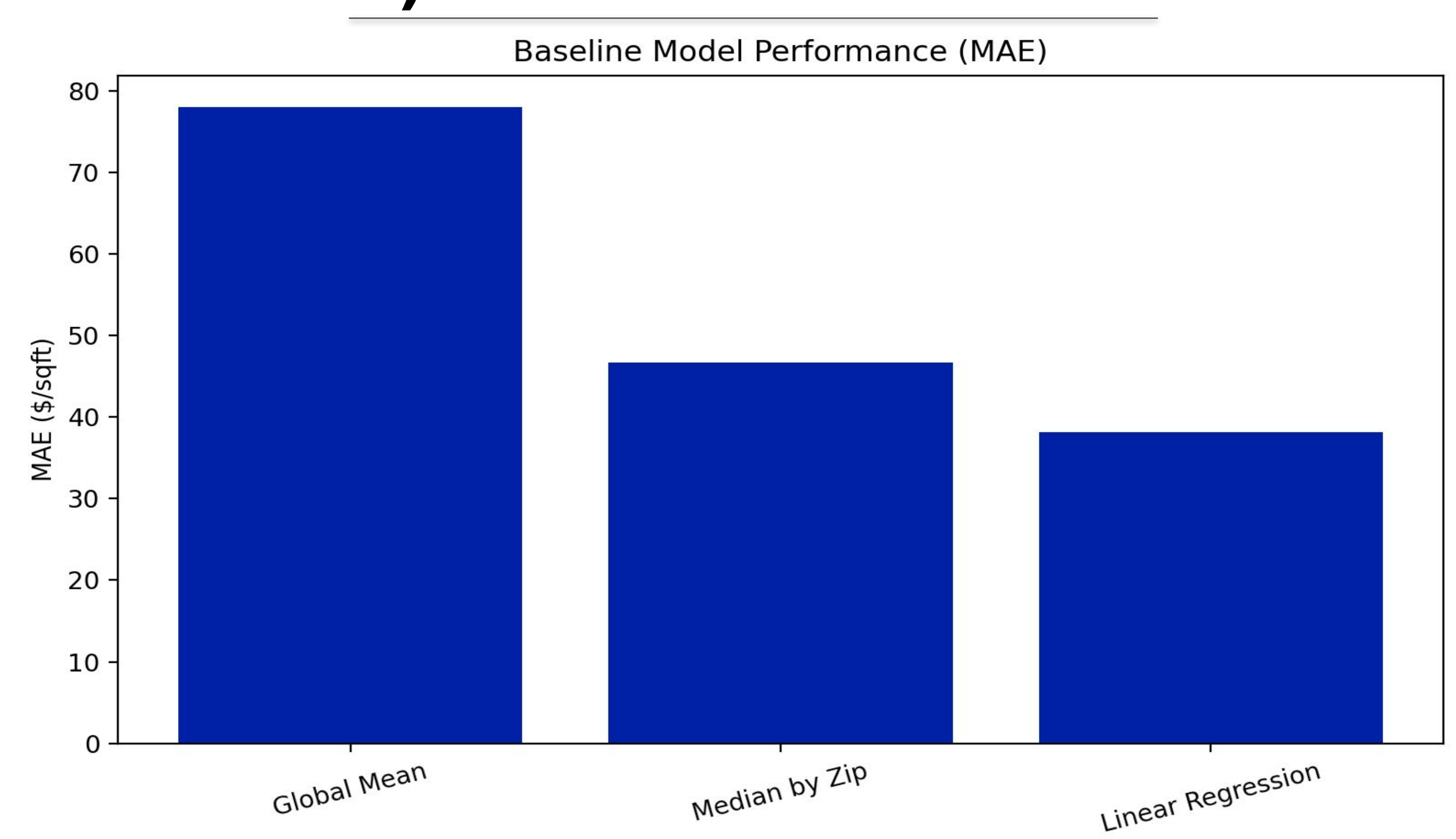
- Trained a **Random Forest Regressor** as the primary ML model
- Tuned number of trees and depth (avoid overfitting)
- Evaluated using **MAE** and **RMSE** on the test set

4) Modeling Approach

Goal: Compare simple interpretable models to a more flexible machine learning model

Baseline models
Global Mean Predictor: establishes a worst case reference
Median-by-Zip: A strong, simple, location-based benchmark
Linear Regression: Interpretable model that tests linear assumptions
Machine Learning Model
Random Forest Regressor: combines many decision trees, captures nonlinear relationships, naturally handles feature interactions, robust to outliers, produces feature importance
Why Random Forest
Housing prices involve nonlinear patterns, and the interaction effects between features (location x year x size) are strong. Tree models do not require heavy normalization and typically provide better predictive accuracy than linear baselines.

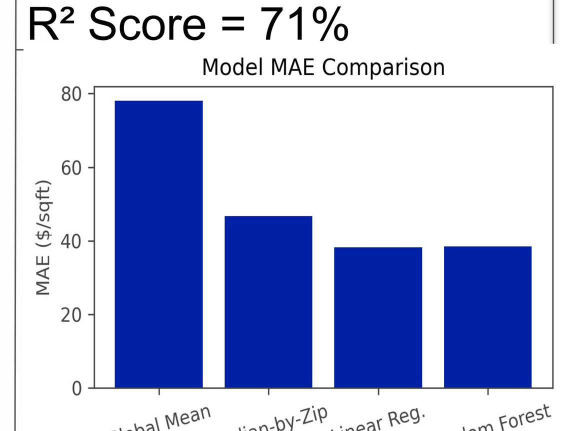
5) Baseline Results



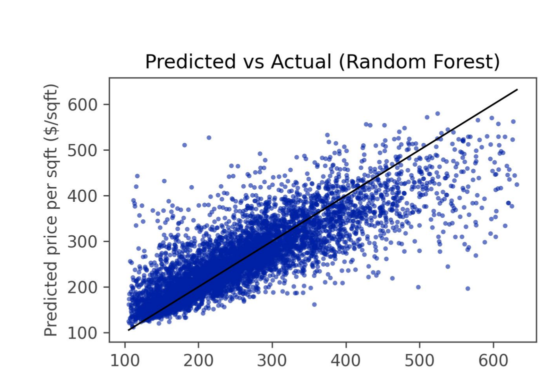
6) Machine Learning Results

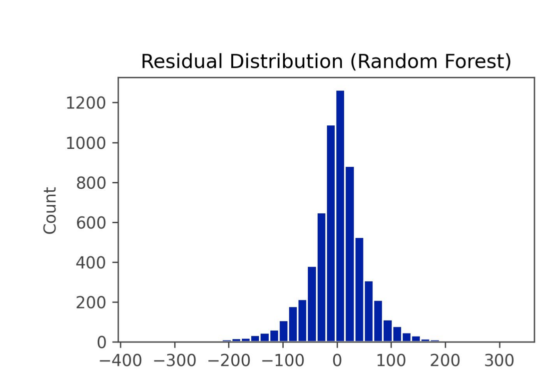
To evaluate model performance beyond baselines, a Random Forest Regressor was trained using all engineered features (numeric + one-hot encoded zip codes).

Test set performance
GMP MAE = 78.0 $/sqft
RMSE = 56.31 $/sqft
MZ MAE = 46.7 $/sqft
LR MAE = 38.4 $/sqft
RanF MAE = 38.4 $/sqft
R^2 Score = 71%



Interpretation
The Random Forest outperforms all baselines. Suggesting strong nonlinear relationships between home attributes and $/sqft.
Location, square footage, and year built play major interactive roles in predicting $/sqft

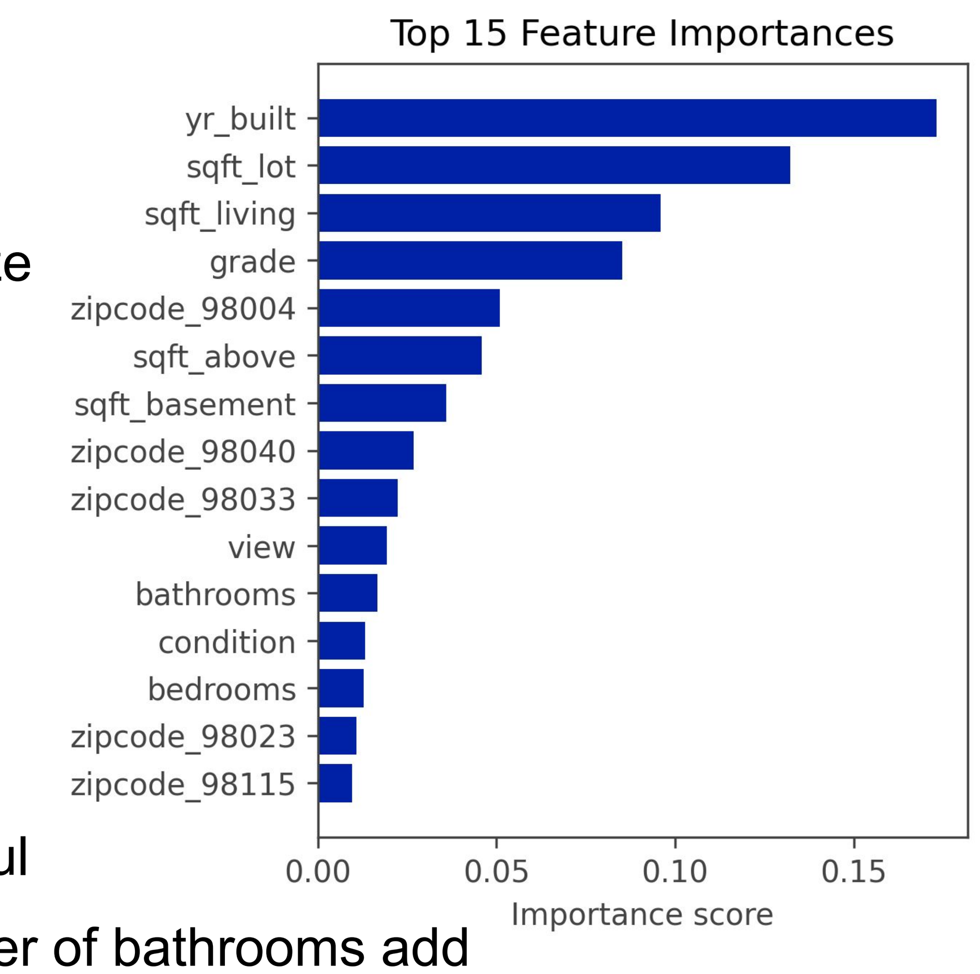




7) Feature Importance Analysis

Top Predictors of $/sqft

- Year built is the strongest predictor
- Square footage and lot size have a large influence, indicating that size efficiency matters
- Home quality (grade) and above-ground sqft contribute heavily
- Location is important but individually less meaningful
- Basement size and number of bathrooms add moderate but meaningful predictive power



8) Key Findings

- Simple models work surprisingly well. Linear regression performed far better than the Global Mean Baseline, showing that basic tabular features explain a large portion of housing price variation
- Random Forest improved accuracy by about 50% when compared to GMP, confirming strong nonlinear relationships in the data
- Location matters but, on average, affect the $/sqft less than others
- Home characteristics (year, square footage, sqft living, grade, and lot size were consistently high-impact drivers of price per sqft
- The model explained 71% of variance in $/sqft ($R^2$ = 0.71), making it strong enough for early-stage neighborhood comparison tasks.
- Feature interactions are critical. The model learned complex patterns (older homes in certain zip codes being very valuable), which the linear models were unable to fully capture

9) Conclusion & Future Work

This project shows that price per square foot can be accurately predicted using only basic features available at listing. A Random Forest Regressor had the best results (MAE = $38.4 / sqft, R^2 = 0.71), outperforming simpler baselines by a wide margin.
The model identified the most critical factors in determining price per square foot as the year it was built, the size of the lot, the livable square footage, and the overall quality of the build. However, location (zip code) proved to be just as interesting as a variable. Most locations had very little effect on the price, but a few zip codes proved to be major factors, whether for good or bad.
In the future, this project would benefit from adding neighborhood features (walkability, school scores, crime, etc.), I think that would significantly alter the importance of location. I would also like to add temporal factors like market trends and seasonality. It would also be interesting to compare different models to see their strengths and faults