King County Housing Price Prediction

Mini IronKaggle Project | Machine Learning Analysis for the Real Estate Market

Aitor Martin | Isis Hassan | Manish Kumar | Suzana Souza

February | 2026

BUSINESS PROBLEM

Project Objectives

Predict Prices

Develop accurate predictive models for sale values

Identify Drivers

Determine the most influential characteristics on price

Premium Segment

Analyze properties above $650K

Informed Decisions

Support data-driven strategies in the real estate sector

DATASET

Data Overview

21

Features

Structural, geographical, and qualitative variables

1

Year

Transactions from 2014-2015

Location: King County, Washington

Target Variable: Sale price (continuous)

Feature Examples

- Living area (sqft_living)
- Quality rating (grade)
- Geographical coordinates
- Waterfront
- Year of construction
- Neighborhood metrics

Exploratory Analysis

Living Area

Strongest correlation: sqft_living demonstrates a positive linear relationship with price

Grade

High grade indicates premium finishes and significantly impacts value

Location

Latitude and longitude reveal geographic patterns of valuation

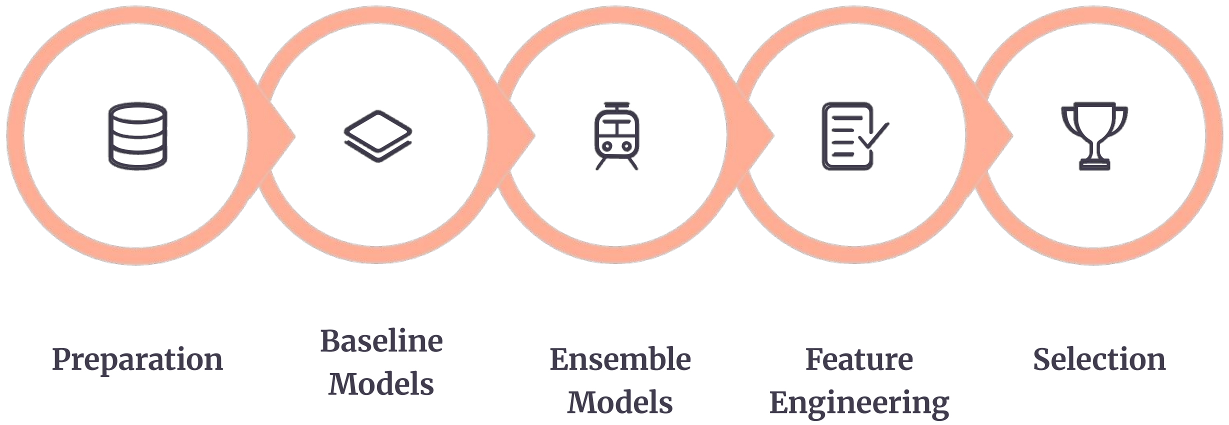
Waterfront

Properties with water access command a substantial premium

We observed non-linear patterns and the presence of outliers, suggesting that simple linear models may not capture the full complexity of the data.

MODELING

Machine Learning Approach



Structured pipeline from pre-processing to final validation

Models Tested

01

Linear Regression

Baseline for comparison

02

ADA

improve the accuracy of weak classifiers

03

KNN

Proximity-based approach

04

Random Forest

Ensemble of decision trees

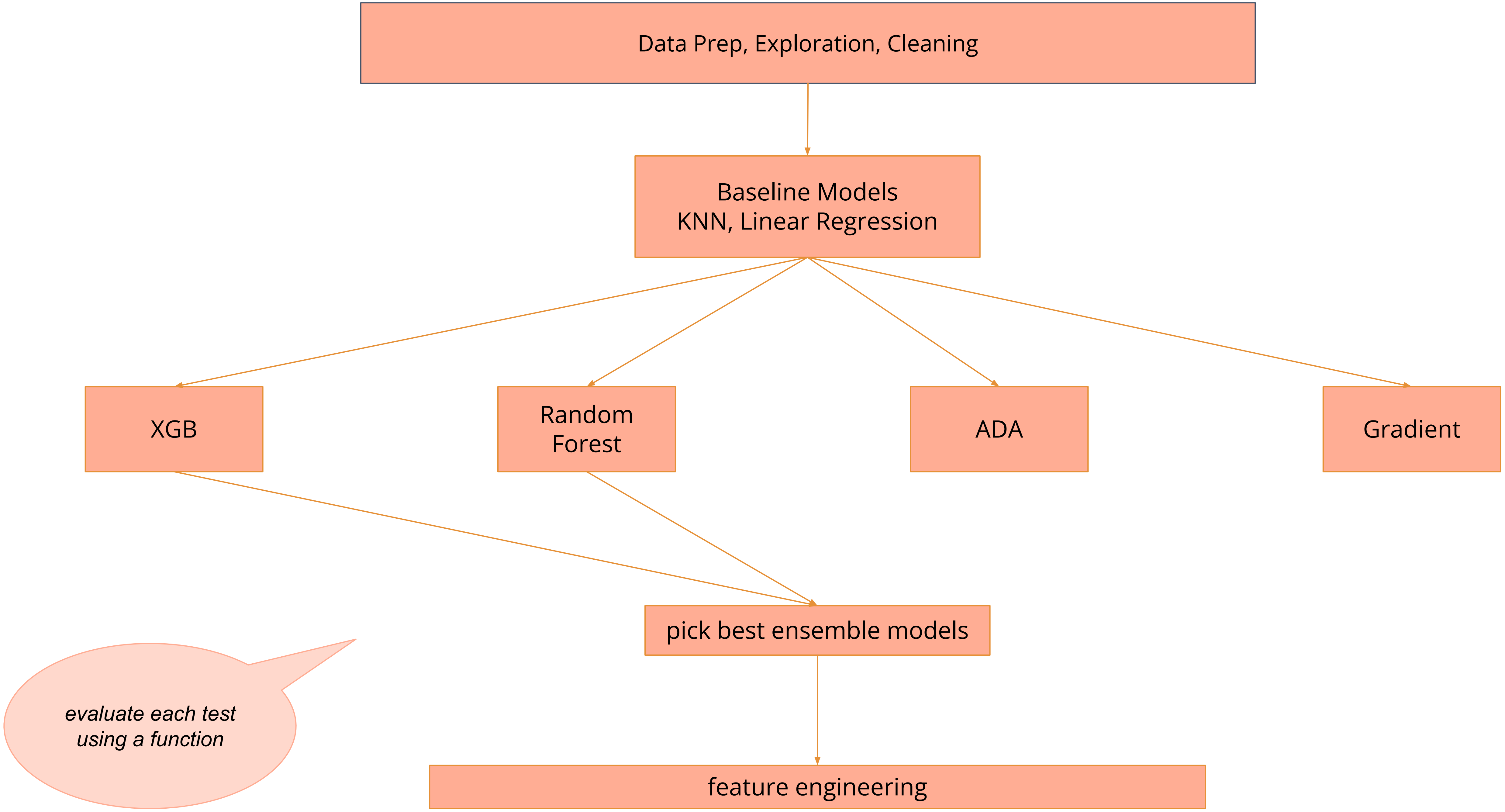
05

Gradient Boosting

Optimized sequential model

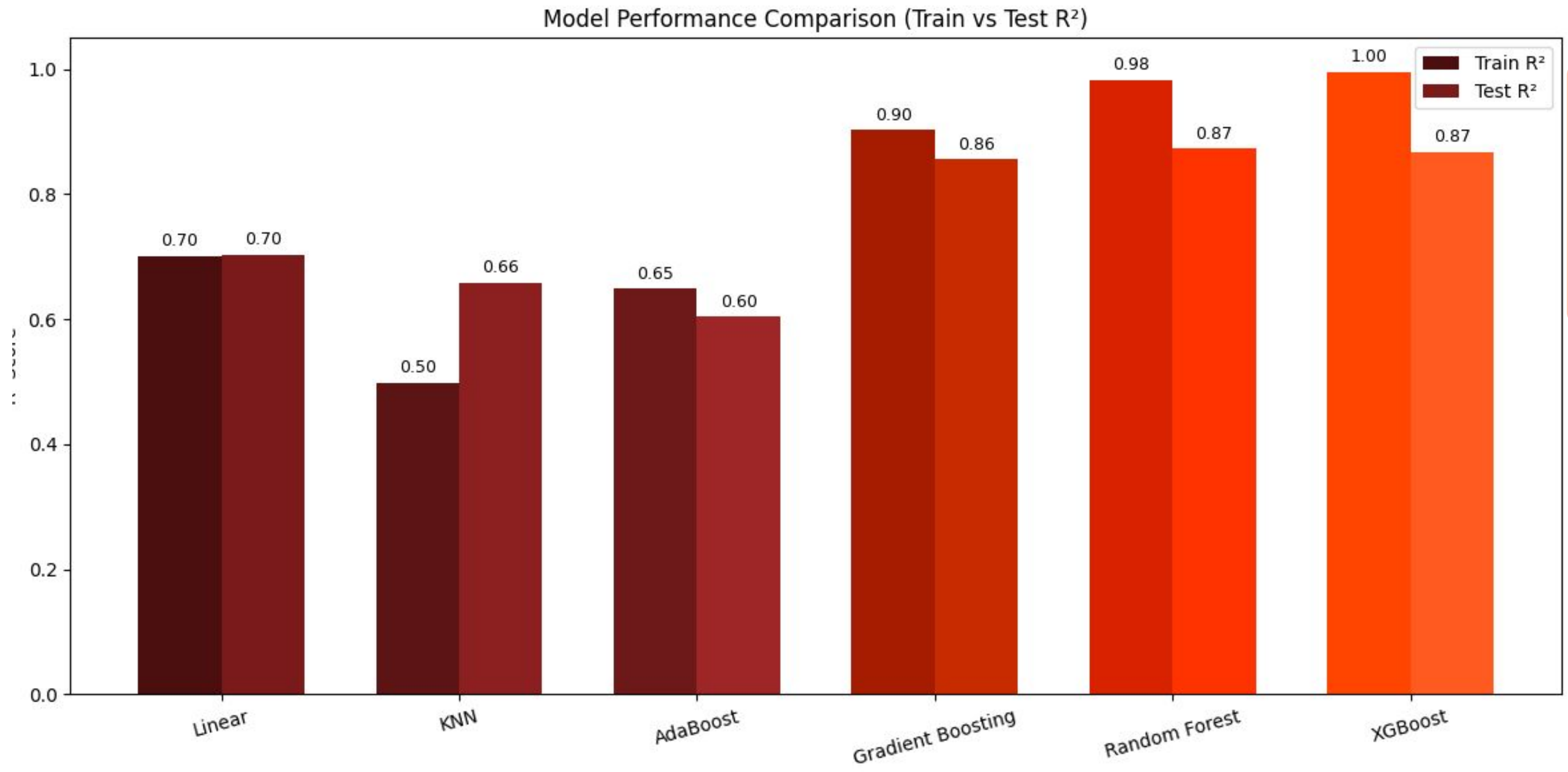
Evaluation Metrics: R^2 (coefficient of determination) and MAE (mean absolute error)

Teamwork


Data Prep, Exploration, Cleaning
Baseline Models
KNN, Linear Regression
XGB
Random Forest
ADA
Gradient
pick best ensemble models
feature engineering
evaluate each test using a function

RESULTS

Model Performance before feature engineeering



✅ **Best Model: XGB**

R^2 = 0.99 | Low MAE

XGBoost and Random Forest performed the best out of all models tested.

Feature Engineering

- Dropped Columns: Sqft_above + sqft_basement = sqft_living, so we can drop redundant sqft columns. Model performance was minimally better when dropping sqft_living", but we still dropped sqft_above + sqft_basement. Why?

    - it removes 2 columns instead of 1
    - it's likelier that real-life data is missing for the subcategories sqft_above and sqft_basement

- Zipcode:
    - one-hot-encoded as a categorical value
- Removed outliers (top/bottom % of sqft_living)
- House Age: Calculated as "Date" - "Yr_Built"
- Yr_Renovated: Turned Boolean (1 = Renovated, 0 = Not Renovated)
- View: Turned (1 = Viewed, 0 = Not Viewed)

Best Model after Feature Engineering

- XGBoost with GridSearchCV
- Feature Engineering: as mentioned on previous slide
- Best Parameters: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 300}

=== Model Evaluation ===
train_r2: 0.9570
test_r2: 0.9071
train_mse: 4300432392.8546
test_mse: 9067113808.9792
train_rmse: 65577.6821
test_rmse: 95221.3937
train_mae: 45672.8331
test_mae: 59339.3464

Q: Can we drop further columns to simplify the model?

A: Probably. We started testing, but so far our tests led to worse model performance.

Next Steps

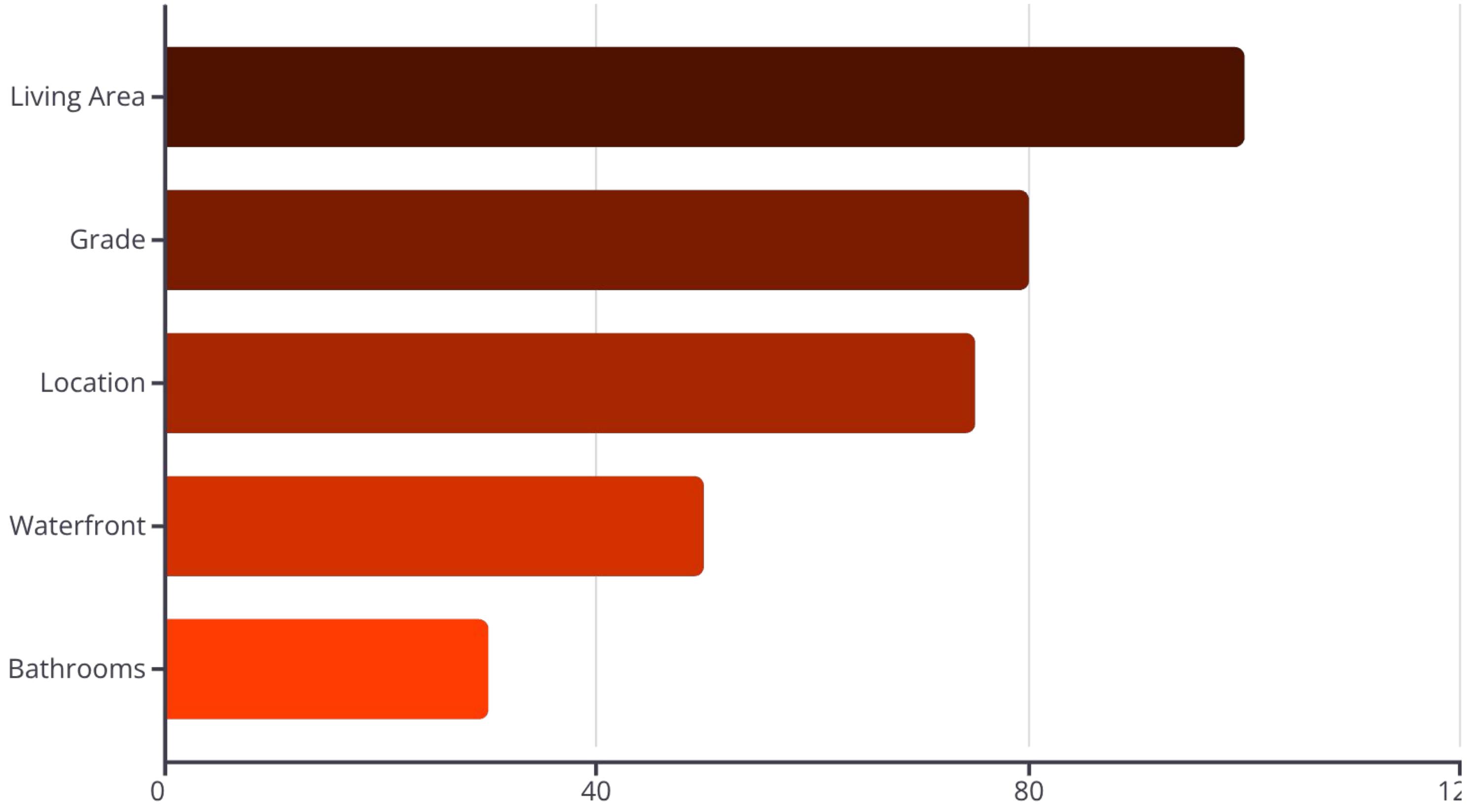
- Further test how dropping columns impacts results
- Test if model stability remains if we drop more less relevant columns
- Cross-validate results

Importance of Features

PRICE DRIVERS

Top 5 Most Influential Features

1. **Living Area (sqft_living)**
   Dominant predictor of market value
2. **Grade**
   Construction quality and finishes
3. **Location (lat/long)**
   Geographic coordinates reveal location premium
4. **Waterfront**
   Water access as a premium differentiator
5. **Bathrooms**
   Indicator of comfort and size



The model confirms that living area and quality grade dominate price determination. Location remains one of the strongest hidden drivers.

High-Value Property Analysis

Property Characteristics > $650K

Larger Areas

Living space substantially above market average

Premium Finishes

Premium finishes and superior quality building materials

Prime Locations

Clustering in high-value geographical zones

Waterfront

Significantly higher prevalence of direct water access

The luxury segment shows distinct clustering around prime location and size. The presence of waterfront access substantially increases value in this segment.

CONCLUSIONS

Key Learnings

Ensemble Superiority

Ensemble models consistently outperform linear approaches in complex real estate data

Location + Size

Two factors dominate real estate price formation

Critical Pre-processing

Adequate data treatment is fundamental for model performance

> This project demonstrates how structured machine learning pipelines generate actionable insights from real estate data, balancing business interpretation and predictive accuracy.

From Data → Insights → Prediction

This project illustrates how machine learning can support real estate pricing decisions through structured data analysis and robust modeling techniques.

Thank you

Questions?