

Decoding Value: King County Housing Analysis

FROM BASELINE BENCHMARKS TO OPTIMIZED PRECISION

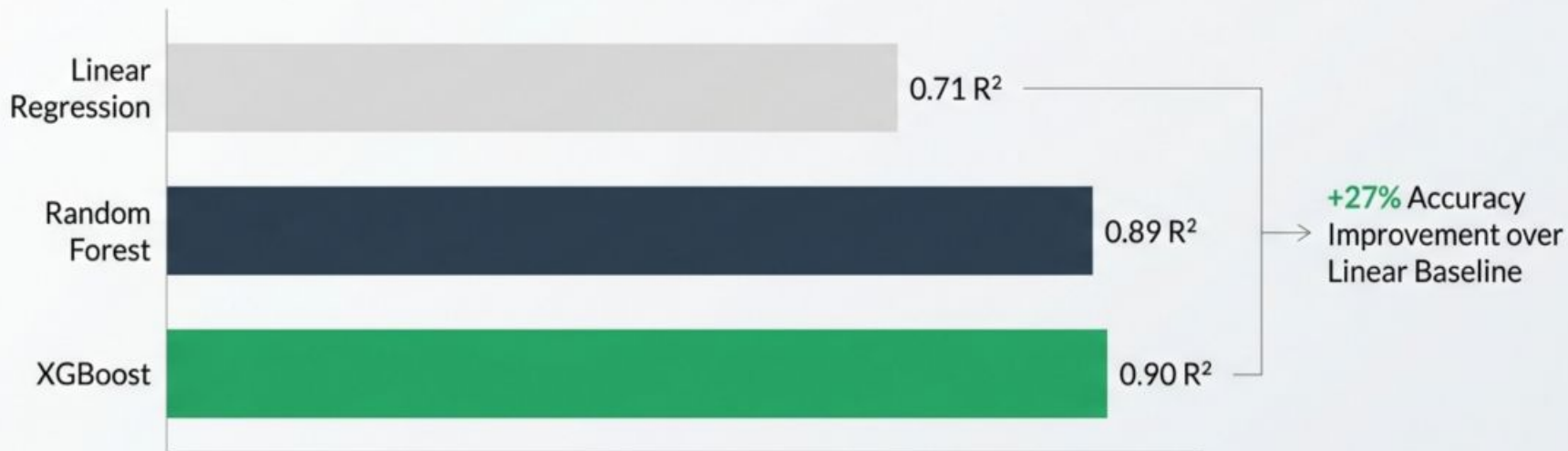
Anne Valvezan
Alejandra Cárdenas
Sheetansh Kaushik



Baseline

Establishing the Baseline

Tree-based ensembles immediately outperform linear approaches.

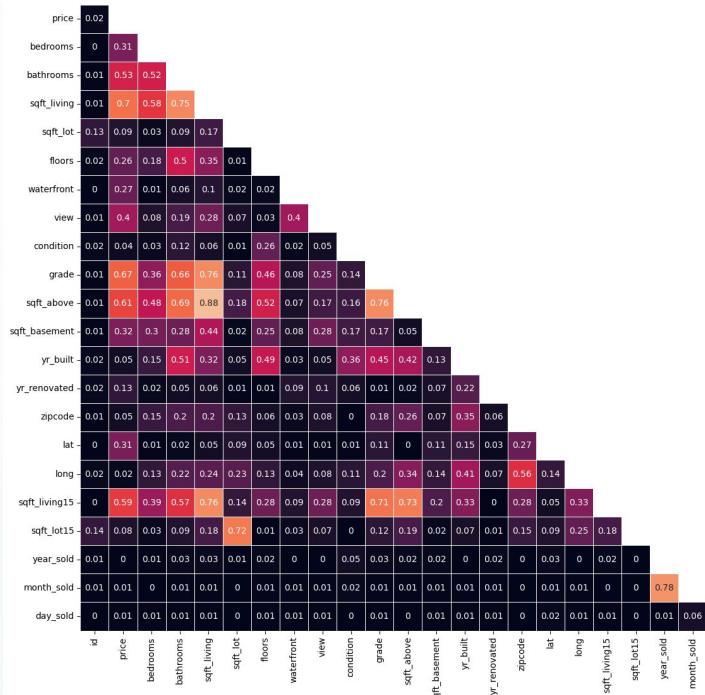


Baseline Models

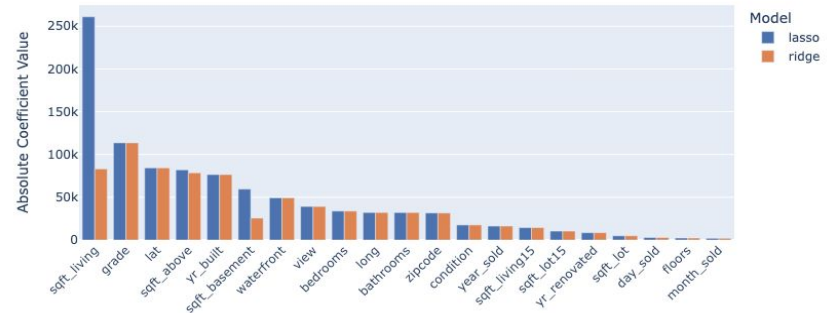
Model	Split	R2	Adjusted_R2	MAE	MAPE	RMSE	Comments
LinearRegression	train	0.6977	0.6974	125948.1118	0.2561	202864.5703	Baseline model
LinearRegression	test	0.7162	0.7148	125985.6747	0.2596	191531.3335	Baseline model
LinearRegression	train	0.6977	0.6974	125948.1118	0.2561	202864.5703	Normalized version of the baseline
LinearRegression	test	0.6834	0.6819	137988.4158	0.2836	202287.1258	Normalized version of the baseline
RandomForestRegressor	train	0.982	0.9819	25948.4444	0.0486	49547.1197	Baseline, no normalization, random_state 13, default values.
RandomForestRegressor	test	0.8959	0.8954	67269.877	0.1271	115994.2051	Baseline, no normalization, random_state 13, default values.
XGBRegressor	train	0.978	0.978	39126.7558	0.0872	54676.7727	Baseline, no normalization, default values.
XGBRegressor	test	0.9015	0.901	65712.7448	0.1246	112860.4995	Baseline, no normalization, default values.

The Drivers of Value

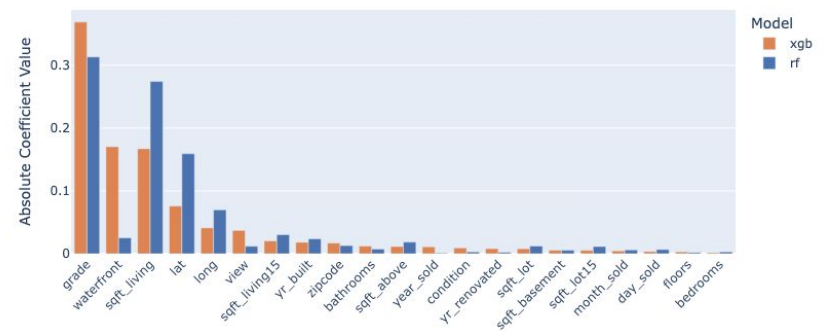
Location, size, and quality dominate the signal across all model types.



Feature Importance: Lasso vs Ridge



Feature Importance: Random Forest vs XGBoost



Trimming the Noise

Removing temporal and structural redundancies had marginal impact on performance.

~~yr_renovated~~

~~sqft_lot~~

~~floors~~

~~day_sold~~

~~month_sold~~

Performance Impact:
<0.01 Change in Test R^2

Model Robustness Confirmed. Dropping these features streamlined the model without sacrificing predictive power.

Contextual Feature Engineering

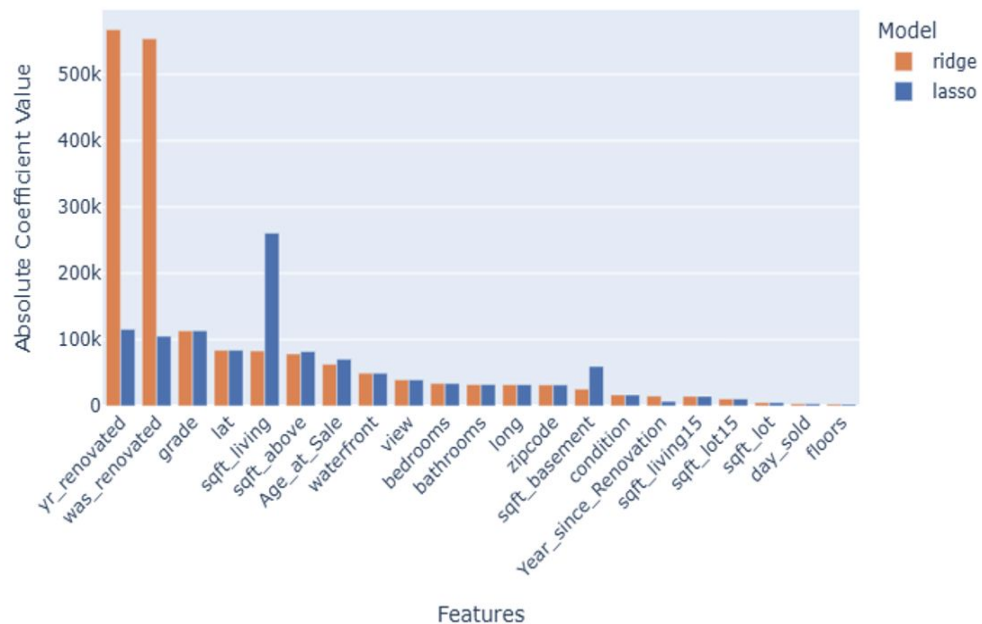
Transforming administrative metadata into human-readable context.



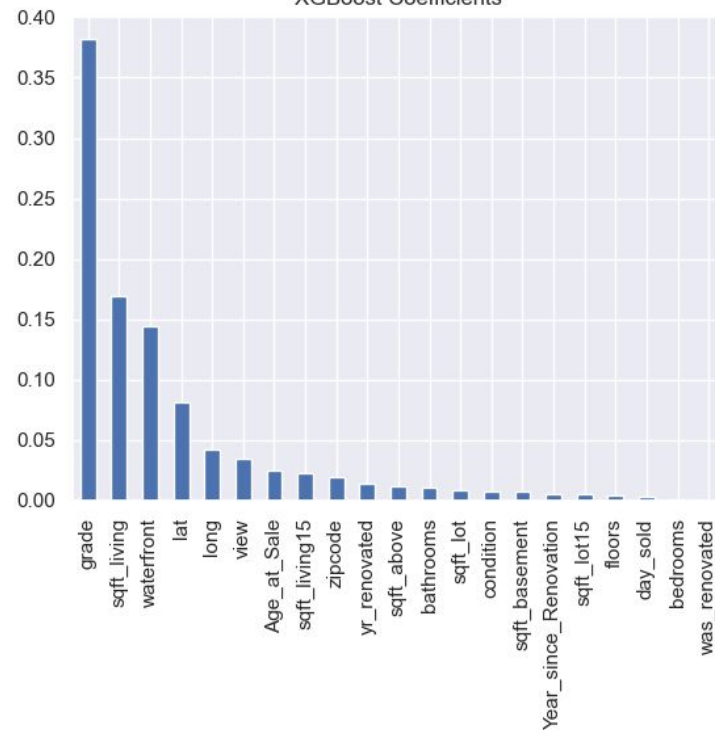
Result: Improved interpretability with stable accuracy (XGBoost $R^2 \sim 0.9038$).

Feature Importance on engineered features

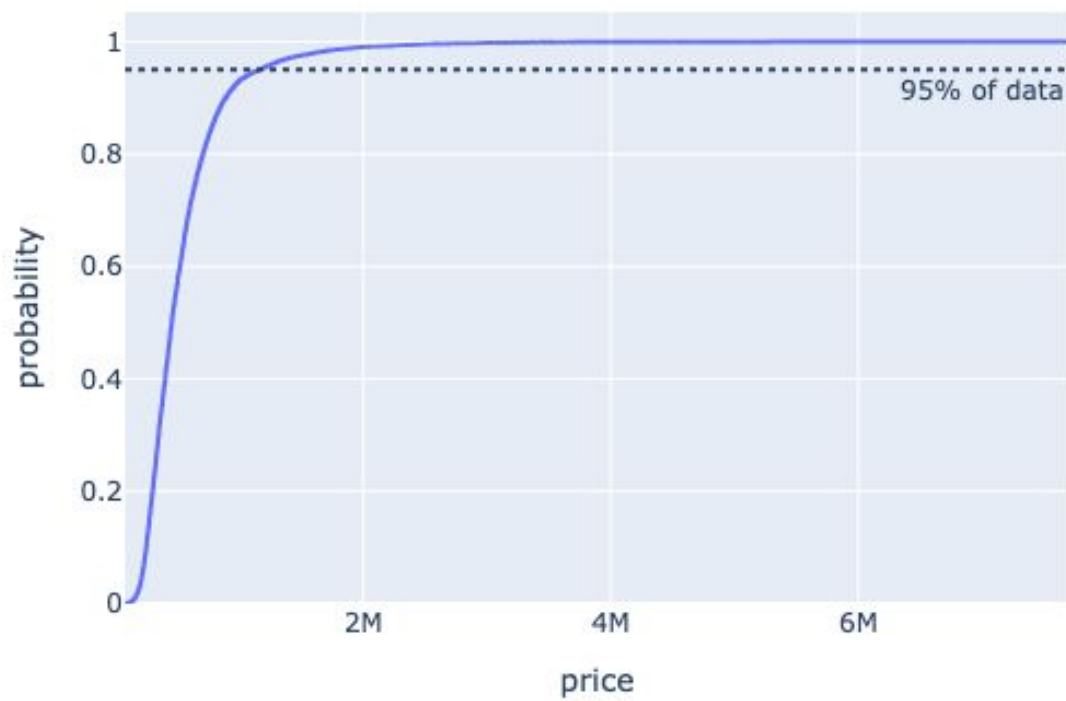
Feature Importance: Lasso vs Ridge



XGBoost Coefficients



Cumulative Distribution of House Prices



Outlier Metrics

Model	Split	R2	Adjusted_R2	MAE	MAPE	RMSE	Comments
RandomForestRegressor	train	0.9879	0.9879	22769.7215	0.0456	40566.2303	Outliers flagging, no normalization.
RandomForestRegressor	test	0.9322	0.9319	59863.8409	0.1203	93599.4998	Outliers flagging, no normalization.
RandomForestRegressor	train	0.9826	0.9826	23185.3777	0.0475	37273.0054	Removing top 1%, no normalization.
RandomForestRegressor	test	0.8739	0.8733	64327.4972	0.129	104088.2866	Removing top 1%, no normalization.
RandomForestRegressor	train	0.9806	0.9806	19934.893	0.0458	29386.2798	Removing top 5%, no normalization.
RandomForestRegressor	test	0.8717	0.871	53038.5069	0.1212	76308.9748	Removing top 5%, no normalization.
XGBRegressor	train	0.9825	0.9825	35584.0027	0.0818	48751.1351	Outliers flagging, no normalization.
XGBRegressor	test	0.926	0.9256	59532.0485	0.1184	97798.9764	Outliers flagging, no normalization.
XGBRegressor	train	0.9679	0.9679	36410.9174	0.083	50584.7096	Removing top 1%, no normalization.
XGBRegressor	test	0.8891	0.8885	61265.8381	0.1234	97618.3705	Removing top 1%, no normalization.
XGBRegressor	train	0.9569	0.9568	32103.1211	0.0783	43810.9507	Removing top 5%, no normalization.
XGBRegressor	test	0.8806	0.88	51602.5277	0.1189	73599.338	Removing top 5%, no normalization.

The Pivot Point: Outlier Strategy

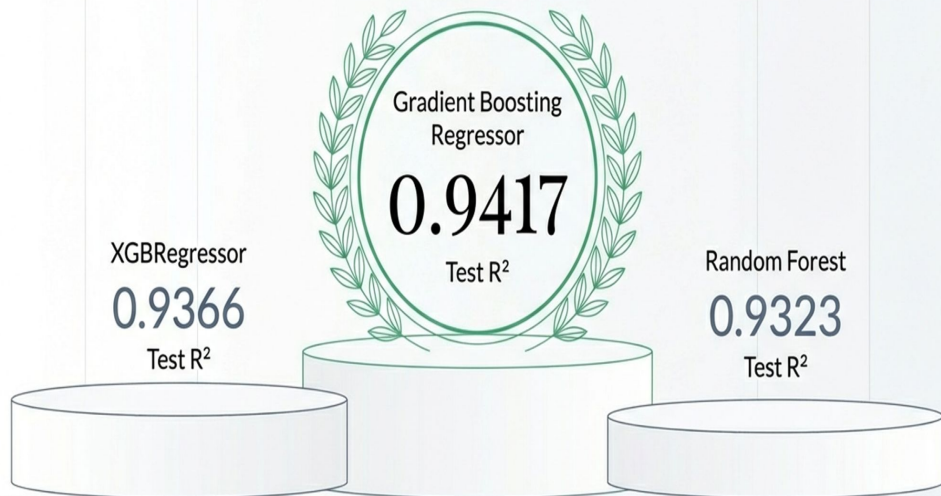
Embracing outliers proved superior to removing them.



The Summit : Gradient Boosting

The Summit: Model Optimization

Gradient Boosting achieves state-of-the-art accuracy.



Hyperparameters for the winning Gradient Boosting Regressor.

```
n_estimators: 500
learning_rate: 0.1
max_depth: 5
min_samples_split: 10
min_samples_leaf: 4
subsample: 0.8
max_features: 'log2'
```

Model Optimization (Grid Search) - V1 (outliers)

Model	Split	R2	Adjusted_R2	MAE	MAPE	RMSE
RandomForestRegressor	train	0.9824	0.9824	24386.7475	0.0499	48930.8643
RandomForestRegressor	test	0.9323	0.9319	59310.3415	0.1198	93538.6406
XGBRegressor	train	0.9724	0.9724	43582.308	0.0964	61287.6651
XGBRegressor	test	0.9366	0.9362	57887.7513	0.1155	90536.7618
GradientBoostingRegressor	train	0.9752	0.9752	41519.4529	0.0925	58051.1787
GradientBoostingRegressor	test	0.9417	0.9414	57844.3576	0.1168	86770.3082

Model Optimization (Grid Search) - V2 (outliers + dropping)

Model	Split	R2	Adjusted_R2	MAE	MAPE	RMSE
RandomForestRegressor	train	0.9924	0.9924	16516.5529	0.0343	32219.5225
RandomForestRegressor	test	0.9355	0.9352	60168.2627	0.1234	91328.2004
XGBRegressor	train	0.9693	0.9693	45770.7327	0.1007	64638.048
XGBRegressor	test	0.9357	0.9355	58589.7346	0.1174	91140.137
GradientBoostingRegressor	train	0.9742	0.9741	42373.1298	0.0948	59316.4161
GradientBoostingRegressor	test	0.9367	0.9365	59336.1257	0.1191	90423.8246

Model Optimization(Grid Search) - V3 (outliers + dropping + feature eng)

Model	Split	R2	Adjusted_R2	MAE	MAPE	RMSE
RandomForestRegressor	train	0.9792	0.9792	29830.8495	0.061	53176.4381
RandomForestRegressor	test	0.9318	0.9315	61500.1162	0.126	93872.8212
XGBRegressor	train	0.9753	0.9752	41858.725	0.0938	58040.5476
XGBRegressor	test	0.9358	0.9355	58910.2638	0.119	91103.133
GradientBoostingRegressor	train	0.9582	0.9581	52551.4223	0.1126	75480.4486
GradientBoostingRegressor	test	0.9347	0.9345	61386.0436	0.1239	91838.4174

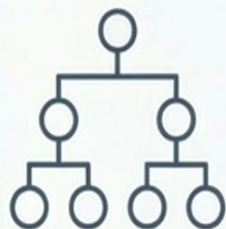
Lessons Learned

Three pillars of predictive improvement.



Handling > Hiding

Outlier flagging retains valuable signal that removal destroys.
High-value homes are part of the market, not errors.



Algorithm Dominance

Boosting methods (Gradient Boosting / XGB) consistently edge out Random Forest and crush Linear Regression.



Tuning Matters

Hyperparameter optimization squeezed the final 1-2% of performance, pushing the R^2 from ~ 0.93 to $0.94+$.