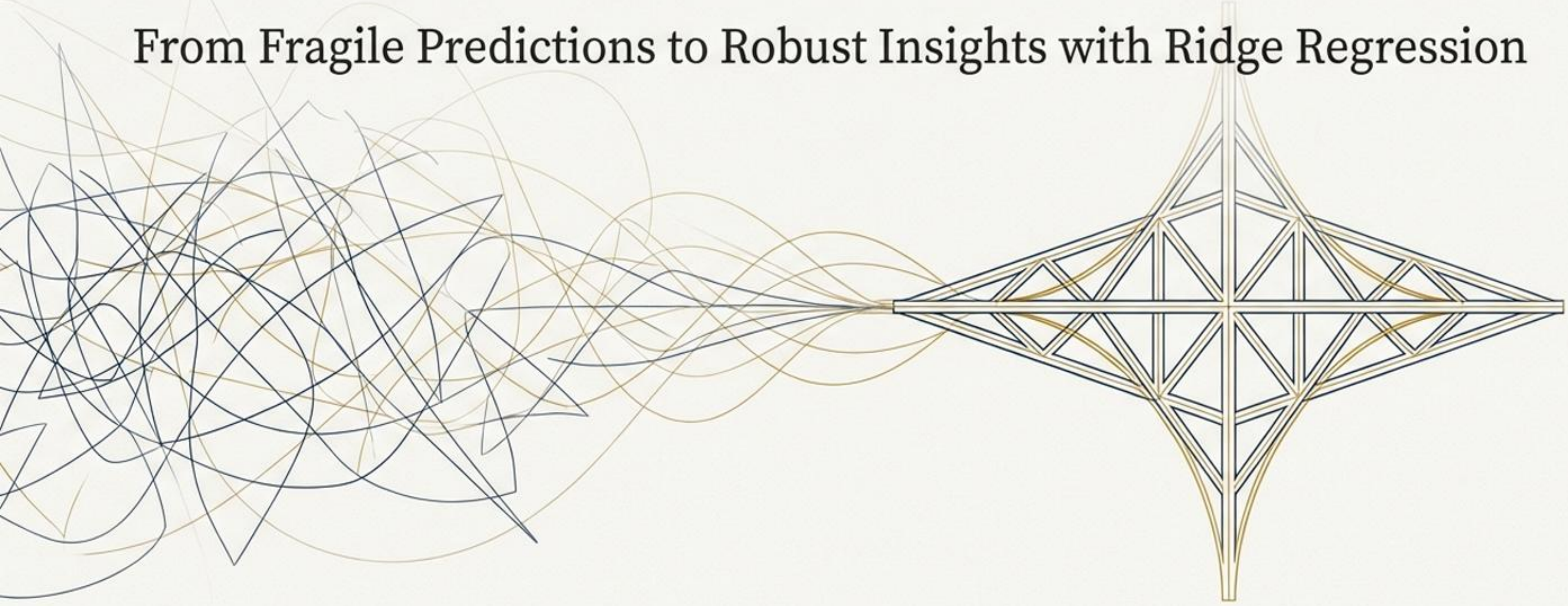
The Quest for a Stable Model

From Fragile Predictions to Robust Insights with Ridge Regression



The Challenge: Predicting Boston Housing Prices

The project goal is to build a predictive model for the median value of owner-occupied homes (`MEDV’) in Boston. We will use a dataset of socio-economic, environmental, and structural attributes to achieve this.

Key Dataset Specs

- **Entries:** 506
- **Features:** 13 predictors, 1 target variable

Key Features to Watch

Target: `MEDV` (Median home value)

Primary Positive Driver: `RM’ (Avg. rooms per dwelling)

Primary Negative Driver: `LSTAT’ (% lower status population)

The Data's Hidden Traps: Uncovering Multicollinearity

A preliminary investigation of the data reveals critical issues that threaten model stability. Standard approaches will fail.

Key Findings

- **Extreme Feature Correlation:** The accessibility to highways (RAD) and property tax rates (TAX`) are almost perfectly correlated.
  - **Correlation Coefficient: 0.91**
- **Inflated Variance:** This relationship was confirmed using the Variance Inflation Factor (VIF).
  - **VIF Scores:** `RAD` and `TAX` both had **scores > 10**, a clear indicator that model coefficients will be unreliable.

The Obvious Suspect: Why Standard Linear Regression Fails

OLS Explained

Ordinary Least Squares (OLS) works by minimizing the sum of squared errors. It's the default, baseline model for regression.

$$J(\theta) = \sum (y^{(i)} - \hat{y}^{(i)})^2$$

The Critical Flaw

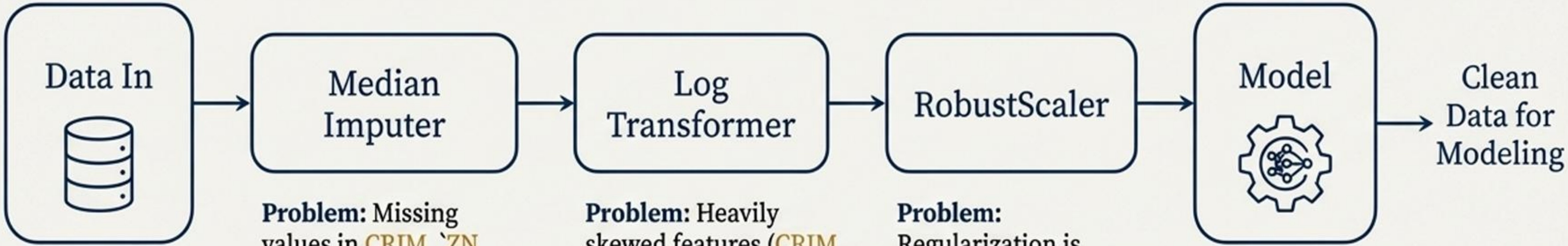
When features are highly correlated, the OLS estimation process becomes unstable.

- **Symptom:** The model produces coefficients with excessively large, opposing values (e.g., +1000 for `RAD`, -995 for `TAX`).
- **Consequence:** The model has High Variance. Tiny changes in the training data cause wild swings in predictions, making the model untrustworthy.

Unstable Predictions

Building a Strong Foundation: The Preprocessing Pipeline

To create a reliable model, we must first clean and transform the data methodically. All steps are wrapped in a Scikit-Learn `Pipeline` to prevent data leakage.



Problem: Missing values in CRIM, `ZN, `AGE, `LSTAT`.

Solution: **Median Imputation**, chosen because it is robust to the extreme outliers present in skewed features like crime rate (CRIM).

Problem: Heavily skewed features (CRIM, `ZN, `DIS) violate linearity assumptions.

Solution: **Log Transformation** (**np.log1p**) to normalize their distributions.

Problem: Regularization is sensitive to the scale of features.

Solution: Use of `**RobustScaler**`.

A Deliberate Choice: Why `RobustScaler` is Essential

For regularization to work correctly, features must be **scaled.**
However, the *type* of scaler is critical when outliers are present.

StandardScaler (The Wrong Tool)

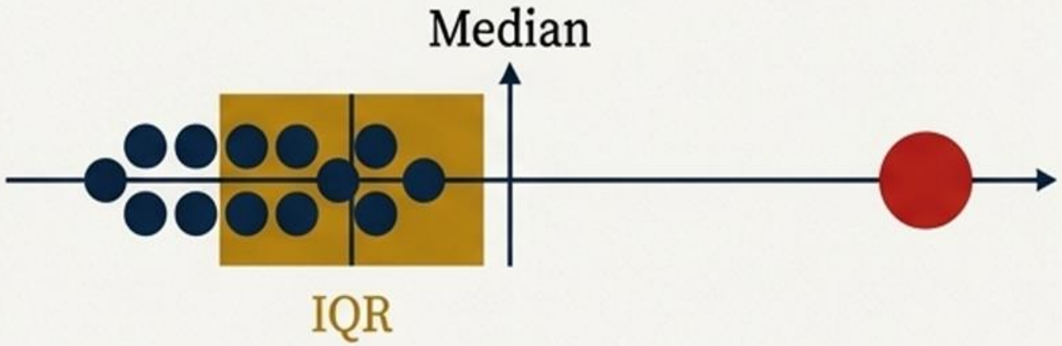
Mechanism: Uses the Mean and Standard Deviation.

Vulnerability: The mean is highly sensitive to outliers. Extreme values (like in the `CRIM` feature) will corrupt the scaling for the entire dataset.

RobustScaler (The Right Tool)

Mechanism: Uses the **Median** and **Interquartile Range (IQR).**

Advantage: Both the median and IQR are highly resistant to outliers. This ensures that the scale of our features remains meaningful and undistorted.



The Solution: Taming Volatility with Ridge Regression

What is Ridge Regression?

Ridge adds a penalty to the OLS loss function, constraining the size of the model's coefficients. This is also known as L2 Regularization.

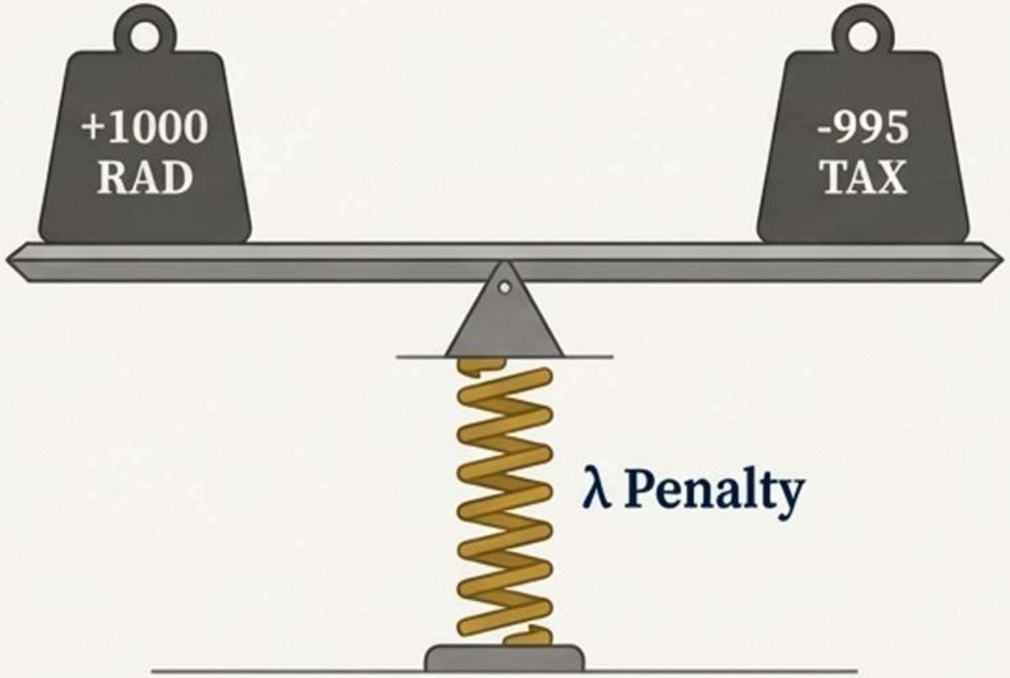
The New Loss Function:

$$J(\theta) = \sum (y^{(i)} - \hat{y}^{(i)})^2 + \lambda \sum \beta_j^2$$

Key Term: Penalizes large weights. λ (alpha) controls the penalty's strength.

The Shrinkage Effect:

- Ridge does not force coefficients to become exactly zero.
- Instead, it **shrinks** the coefficients of correlated predictors towards each other and towards zero. This dampens the noise from multicollinearity and stabilizes the model.



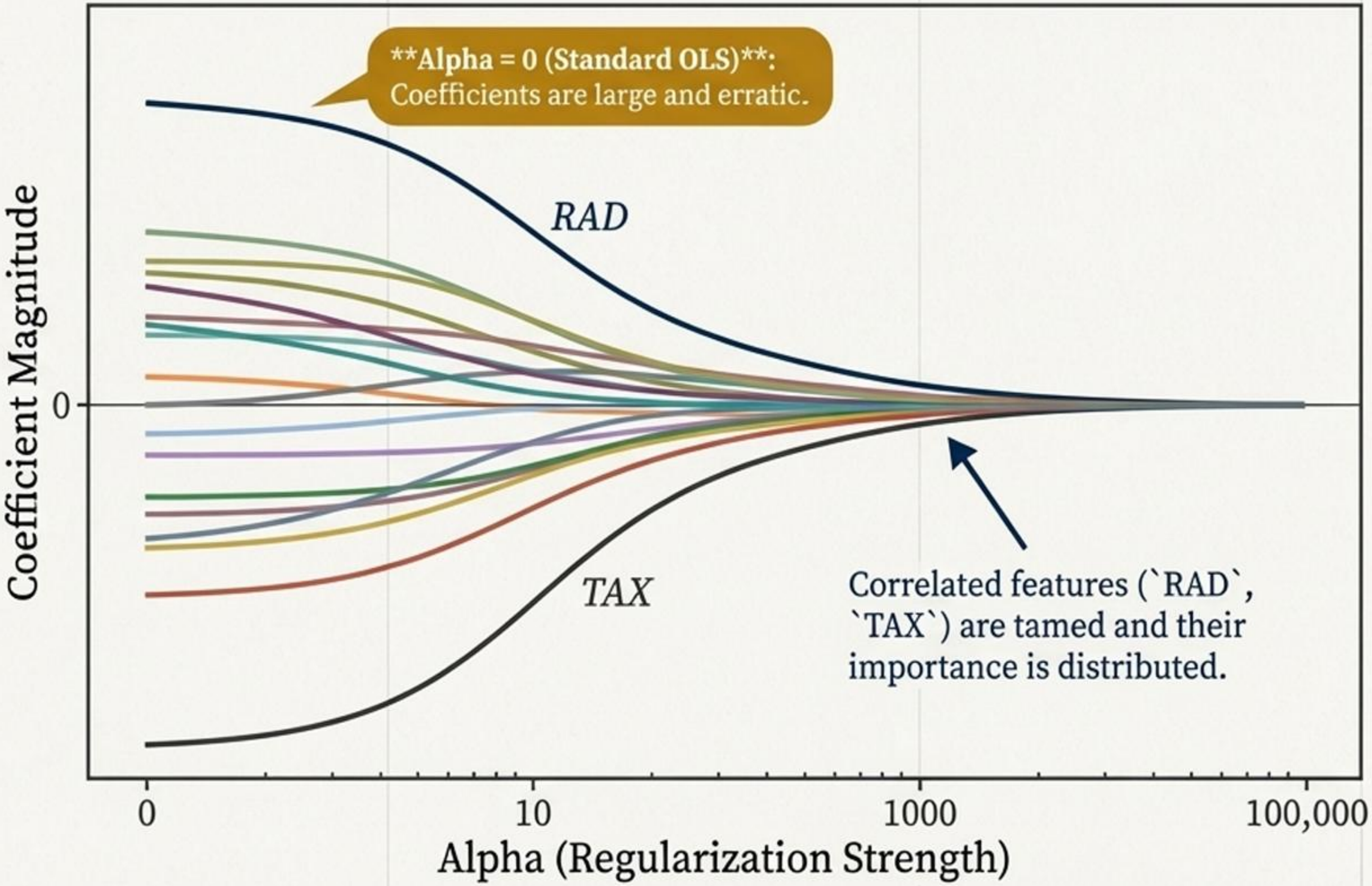
Stable Predictions

Visualizing the Shrinkage: The Ridge Trace Plot

What this Chart Shows: This plot displays how the magnitude of each feature's coefficient changes as we increase the regularization strength (Alpha / λ).

Key Observations:

- At Alpha = 0, the model is standard OLS, and coefficients are large and erratic.
- As Alpha increases, the coefficients are "tamed," shrinking smoothly towards zero.
- Notice how the coefficients for the highly correlated features (`RAD`, `TAX`) converge, demonstrating how Ridge distributes their importance.



On the Surface, Performance Appears Identical

Methodology: Models were trained and then evaluated on an unseen 20% test set.

Performance Metrics Comparison

Metric	OLS (Linear Regression)	Ridge Regression
RMSE	4.19	**4.20**
R^2 Score	0.760	**0.758**

The Question: If the scores are the same, what was the benefit? The answer lies not in overall accuracy, but in stability and reliability.

The Real Test: Stability Under Pressure

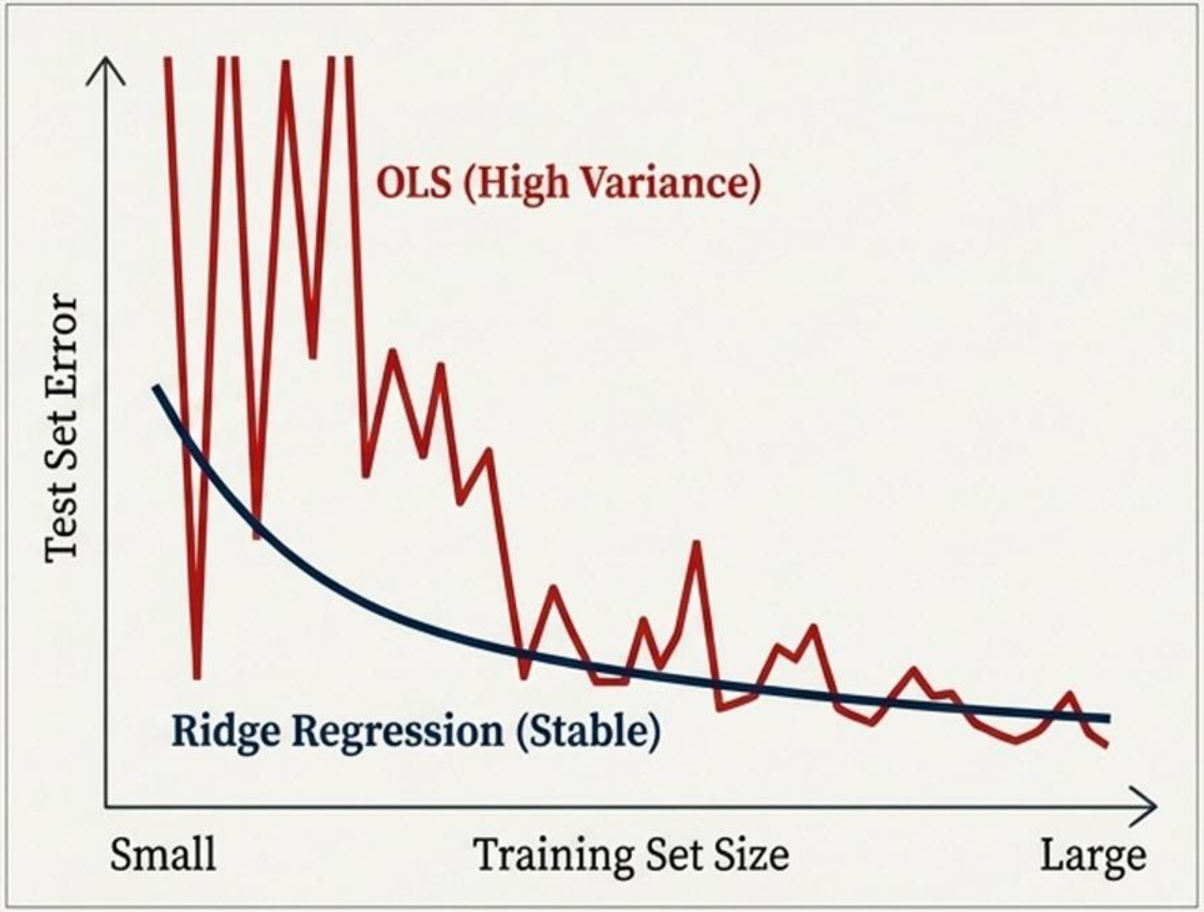
The Insight: Headline metrics on a full dataset can be misleading. A model's true value is revealed when data is scarce or noisy.

The Simulation:

We re-trained both models on a tiny, random subset of only **30 samples** to simulate a data-scarce scenario.

The Result:

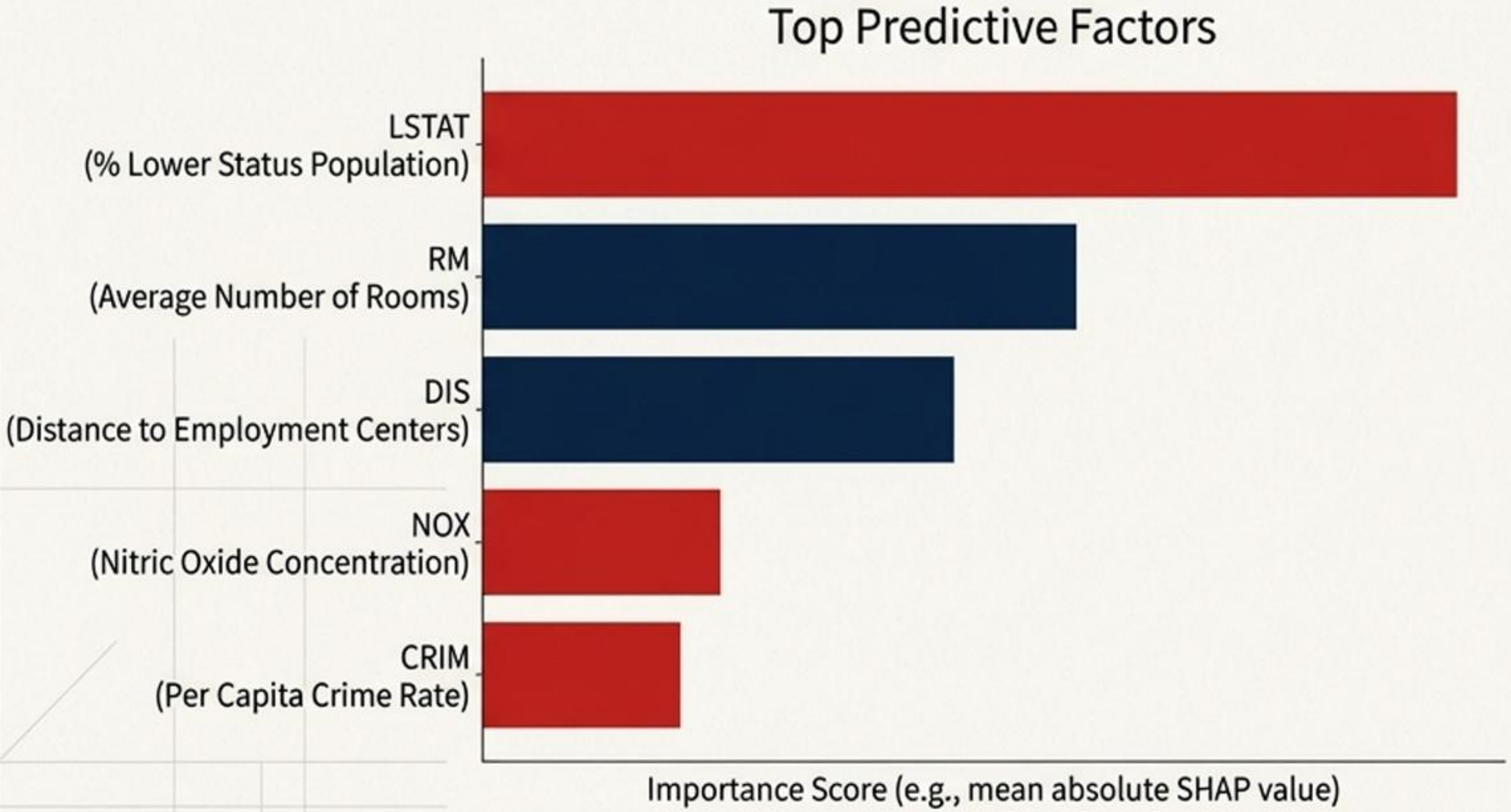
- **OLS (High Variance):** Overfit massively to the small dataset, resulting in poor, unreliable predictions on the test set.
- **Ridge Regression (Stable):** The regularization penalty prevented overfitting. Ridge maintained reasonable error rates, proving its robustness.



Conclusion: **Ridge Regression produces a mathematically 'safer' and more reliable model, especially for future data.**

What Drives Housing Prices? Interpreting the Model

Methodology: Using SHAP (SHapley Additive exPlanations) and Permutation Importance, we identified the most impactful features in the final Ridge model.



1. **LSTAT** (% Lower Status Population)
Impact: Strongly Negative. The single most powerful predictor of lower housing prices.

2. **RM** (Average Number of Rooms)
Impact: Strongly Positive. More rooms consistently lead to higher prices.

3. **DIS** (Distance to Employment Centers)
Impact: Positive. Greater distance from industrial zones is associated with higher home values.

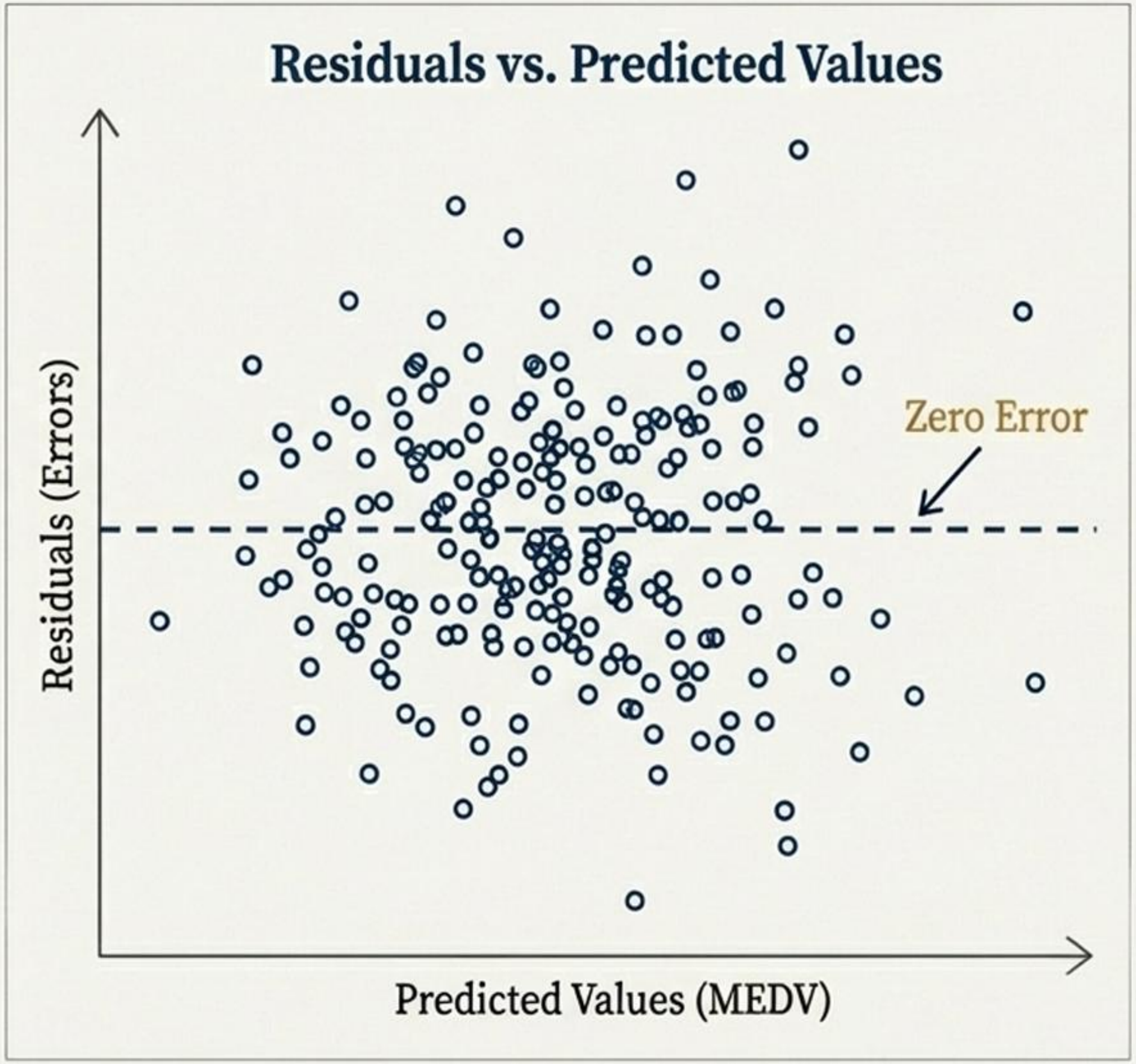
Final Diagnostic: A Clean Bill of Health

The Test: A residual plot graphs the model's prediction errors against the predicted values. A healthy model should show no discernible pattern.

Our Model's Result: The plot shows a random cloud of points centered around the zero-error line.

What This Confirms:

- **Homoscedasticity:** The variance of the errors is constant.
- **No Hidden Bias:** The model has successfully captured the primary linear relationships in the data.
- **Reliable Predictions:** There are no systematic errors; the model is not, for example, consistently underpredicting high-value homes.



From Analysis to Actionable Principles

1. Preprocessing is as Critical as the Algorithm.

The careful selection of **RobustScaler** and the use of **Log Transformations** were essential for model success. Don't just focus on the final algorithm.

2. Prioritize Stability, Not Just a Score.

Ridge Regression provides stability and reliability. It's a 'safer' model that is more likely to perform well on future, unseen data, even if its R^2 isn't dramatically higher.

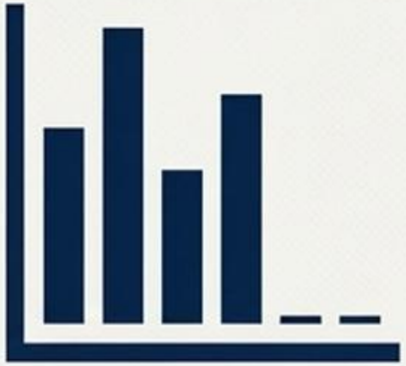
3. Enforce Integrity with Pipelines.

Using Scikit-Learn's **Pipeline** framework is a non-negotiable best practice. It guarantees no data leakage and ensures that results are valid and reproducible.

Future Investigations: Building on This Foundation

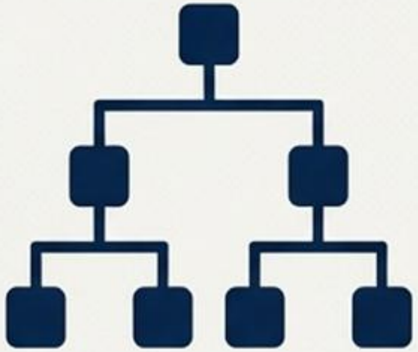
While the Ridge model is robust for linear relationships, further enhancements are possible.

Automated Feature Selection with ElasticNet



Explore ElasticNet regression, a hybrid that combines the coefficient shrinkage of Ridge with the feature-selection capability of Lasso (L1). This could simplify the model by removing less important features.

Capturing Non-Linearity with Tree-Based Models



Implement models like Random Forest or Gradient Boosting to capture complex, non-linear interactions between features that Ridge, by its nature, cannot.

THANK YOU