Project Report: Predicting Housing Prices using Ridge Regression

Author: [Sudipta Biswas]

Date: [02-12-2025]

Subject: Machine Learning / Regularization Techniques

---

1. Executive Summary

This project focuses on building a robust predictive model for housing prices using the Boston Housing dataset. The primary challenge with this dataset is the presence of **multicollinearity** (highly correlated features) and **skewed distributions** in key variables like crime rates.

While a standard Linear Regression (OLS) model provides a baseline, it is prone to high variance and instability in the presence of correlated predictors. To mitigate this, **Ridge Regression (L2 Regularization)** was implemented.

The project workflow included extensive Exploratory Data Analysis (EDA), advanced preprocessing (Median Imputation, Log Transformations, and Robust Scaling), and rigorous model diagnostics (Validation Curves, Coefficient Trace Plots, and SHAP Analysis). The final Ridge model demonstrated superior stability and robustness compared to the OLS baseline, particularly in simulated data-scarce scenarios.

---

2. Problem Statement & Dataset Overview

2.1 The Goal

The objective is to predict the median value of owner-occupied homes (MEDV) in $1000s based on various socio-economic, environmental, and structural attributes of towns in Boston.

2.2 The Dataset

The dataset consists of **506 entries** and **14 columns** (13 features + 1 target).

Key Features:

- **Target:** MEDV (Median value of owner-occupied homes).
- **Structural:** RM (Average number of rooms). Strongest positive correlation with price.
- **Socio-Economic:** LSTAT (% Lower status of the population). Strongest negative correlation with price.

- **Environmental:** NOX (Nitric oxide concentration), CHAS (Charles River dummy variable).
- **Problematic Features:** RAD (Highway accessibility) and TAX (Property tax rate) exhibited a correlation of **0.91**, introducing significant multicollinearity.

---

3. Theoretical Background: Ridge Regression

3.1 Why Linear Regression (OLS) Fails

Ordinary Least Squares (OLS) minimizes the standard Sum of Squared Errors (SSE):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

When features are highly correlated (multicollinearity), the matrix inversion required for OLS becomes unstable. This results in coefficients that are excessively large (e.g., $+1000$ for one feature and -995 for another), causing **High Variance**. Small changes in the input data lead to massive swings in predictions.

3.2 The Ridge Solution (L2 Regularization)

Ridge Regression adds a penalty term to the loss function to constrain the size of the coefficients:

$$J(\theta) = \sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})^2 + \lambda \sum_{j=1}^{n} \beta_j^2$$

- λ **(Alpha):** The tuning parameter.
  - If $\lambda = 0$: The model becomes OLS.
  - If $\lambda \to \infty$: The coefficients are forced toward zero (High Bias).
- $\sum \beta_j^2$ **:** This term penalizes large weights.

Key Mechanism: Ridge Regression does not set coefficients to exactly zero (unlike Lasso), but it **shrinks** them. This allows the model to retain all features while dampening the noise caused by multicollinearity.

---

4. Methodology & Implementation

4.1 Exploratory Data Analysis (EDA)

Before modeling, we diagnosed the "health" of the data:

- **Distribution Check:** CRIM (Crime Rate) and ZN (Zoning) were found to be heavily right-skewed. The target MEDV showed a ceiling effect at $50,000 (capped data).
- **Multicollinearity Detection:**
    - **Correlation Heatmap:** Revealed a 0.91 correlation between RAD and TAX.
    - **VIF (Variance Inflation Factor):** RAD and TAX had VIF scores > 10, confirming that OLS estimates would be unreliable.

4.2 Data Cleaning & Imputation

The dataset contained missing values in CRIM, ZN, INDUS, AGE, and LSTAT.

- **Strategy: Median Imputation**.
- **Justification:** Since predictors like Crime Rate were heavily skewed with massive outliers, using the Mean would have biased the imputation. The Median is robust to outliers.

4.3 Feature Engineering & Transformation

1. **Log Transformation:** We applied np.log1p to skewed features (CRIM, ZN, DIS). This transformed their distributions from "long-tailed" to "normal-like," which satisfies the linearity assumption of Ridge Regression.
2. **Interaction Terms:** A new feature LSTAT_RM (LSTAT * RM) was created to capture the combined effect of poverty and house size.

4.4 Preprocessing: The Choice of Scaler

Scaling is **mandatory** for Ridge Regression because the penalty term $\lambda \beta^2$ is sensitive to the magnitude of the data.

- **Selected Scaler: RobustScaler**.
- **Reasoning:** Unlike StandardScaler (which uses Mean/Variance and is corrupted by outliers), RobustScaler scales data using the **Median** and **IQR (Interquartile Range)**. This ensured that the extreme crime outliers did not distort the scale of the entire dataset.

4.5 Model Pipeline

To prevent **Data Leakage**, all steps were encapsulated in a Scikit-Learn Pipeline.

1. **Imputer:** SimpleImputer (Median)
2. **Transformer:** Log Transformation Function
3. **Scaler:** RobustScaler
4. **Model:** RidgeCV (Built-in Cross-Validation to find the optimal Alpha).

---

5. Diagnostics & Visualizations

To ensure the model was not treated as a "black box," several diagnostic plots were generated:

1. **Validation Curve:** We plotted RMSE against a range of Alpha values. The curve showed an L-shape, indicating that while the base model was stable, regularization prevented overfitting as complexity increased.
2. **Ridge Trace Plot:** A plot of Coefficient Magnitude vs. Alpha. As Alpha increased, we observed the coefficients of correlated features (RAD, NOX) converging toward zero, visually proving the "Shrinkage" effect.
3. **Residual Analysis:** A scatter plot of Residuals vs. Predicted Values showed a random cloud of points centered at zero, confirming that the model captured the main linear trends and that errors were homoscedastic (constant variance).

---

6. Results & Discussion

6.1 Performance Metrics

The models were evaluated on an unseen Test Set (20% split).

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

6.2 The Stability Test

While the metrics on the full dataset were similar, the true power of Ridge was revealed in stability tests:

- **Coefficient Analysis:** OLS assigned large, opposing weights to RAD and TAX. Ridge shrank these weights significantly, distributing the predictive power more evenly.
- **Small Data Simulation:** When trained on a subset of only 30 samples:
    - OLS Overfitted massively (High Variance).
    - Ridge maintained reasonable error rates.

6.3 Feature Importance (Interpretation)

Using **Permutation Importance** and **SHAP (SHapley Additive exPlanations)**, the top drivers of housing prices were identified as:

1. **LSTAT:** The percentage of lower-status population (Negative impact).
2. **RM:** Number of rooms (Positive impact).
3. **DIS:** Distance to employment centers (Positive impact).

---

7. Conclusion

This project demonstrated that while Ordinary Least Squares is a powerful tool, it is fragile in the face of multicollinearity and outliers. By implementing **Ridge Regression**, we achieved a model that is mathematically "safer."

Key Takeaways:

1. **Preprocessing Matters:** The use of RobustScaler and Log Transformations was just as critical as the model choice itself.
2. **Ridge = Stability:** Even if Ridge does not drastically lower the RMSE on abundant data, it stabilizes coefficients, making the model more robust to future data drift.
3. **Pipeline Integrity:** Using Pipelines ensured zero data leakage, validating that our results are reliable for real-world deployment.

Future Work:

Further improvements could be made by exploring ElasticNet (combining Ridge and Lasso) to perform feature selection alongside regularization, or by using Tree-based models (Random Forest) to capture non-linearities that Ridge misses.

---

8. References & Tools Used

- **Language:** Python 3.x
- **Libraries:** Pandas, NumPy, Scikit-Learn, Matplotlib, Seaborn, SHAP.
- **Dataset:** UCI Machine Learning Repository (Boston Housing).