NC STATE UNIVERSITY

North Carolina State University

Department of Financial Mathematics

FIM590 003 Machine Learning in Finance

Iowa House Price Prediction with Linear Regression

Author:

Jiayuan Li

Shixun Wu

Siyuan Lu

October 07, 2025

Contents:

FIM 590-003 Machine Learning Project

Iowa House Price Prediction

1. Introduction

In this project, we aim to predict house prices using modern machine-learning techniques. Our objective is to build a regression model with well-chosen features that minimize mean squared error. Following *Machine Learning in Business: An Introduction to the World of Data Science* (Hull, 2021), we compared Linear, Ridge, and Lasso regression models and evaluate their performance on a held-out test set. After that, we used real-world housing data from Zillow.com to ground the analysis.

The dataset contains Iowa home sales. After cleaning, we randomly split the data into training, validation, and test sets. The training and validation sets are used to fit models and pick the best model; the test set is reserved for final evaluation. Our feature set includes 47 variables - both numerical and categorical - capturing characteristics such as lot and living area, structural attributes, and neighborhood indicators.

For Ridge and Lasso, we systematically vary the penalty parameter (λ) and select the model that achieves the lowest validation mean squared error. We then report test-set performance for the selected model to assess generalization.

2. Data Preparation

2.1 Features

The original dataset contains 80 features and 1 target. We construct a feature set of **47** variables comprising 21 numerical features and 26 categorical indicators (1 ordinal, 25 neighborhood dummies):

Part I (21 numerical):

`LotArea`, `OverallQual`, `OverallCond`, `YearBuilt`, `YearRemodAdd`, `BsmtFinSF` (= `BsmtFinSF1` + `BsmtFinSF2`), `BsmtUnfSF`, `TotalBsmtSF`, `1stFlrSF`, `2ndFlrSF`, `GrLivArea`, `FullBath`, `HalfBath`, `BedroomAbvGr`, `TotRmsAbvGrd`, `Fireplaces`, `GarageCars`, `GarageArea`, `WoodDeckSF`, `OpenPorchSF`, `EnclosedPorch`.

These variables are measured in their natural units (e.g., square feet for areas) and later normalized (Section 2.4).

Part II (1 ordinal variable):

`BsmtQual`

We treat this as an ordinal measure of ceiling height/quality using the conventional categories $\{EX, GD, TA, FA, PO, NA\}$ and map them to integers $\{5, 4, 3, 2, 1, 0\}$, respectively.

Part III (25 neighborhood indicators):

`Neighborhood` encodes location. Because location is a major price determinant, we introduce 25 binary indicators:

`NBHD_Blmngtn, NBHD_Blueste, NBHD_BrDale, NBHD_BrkSide, NBHD_ClearCr, NBHD_CollgCr, NBHD_Crawfor, NBHD_Edwards, NBHD_Gilbert, NBHD_IDOTR, NBHD_MeadowV, NBHD_Mitchel, NBHD_NAmes, NBHD_NPkVill, NBHD_NWAm, NBHD_NoRidge, NBHD_NridgHt, NBHD_OldTown, NBHD_SWISU, NBHD_Sawye, NBHD_SawyerW, NBHD_Somerst, NBHD_StoneBr, NBHD_Timber, NBHD_Veenke`.

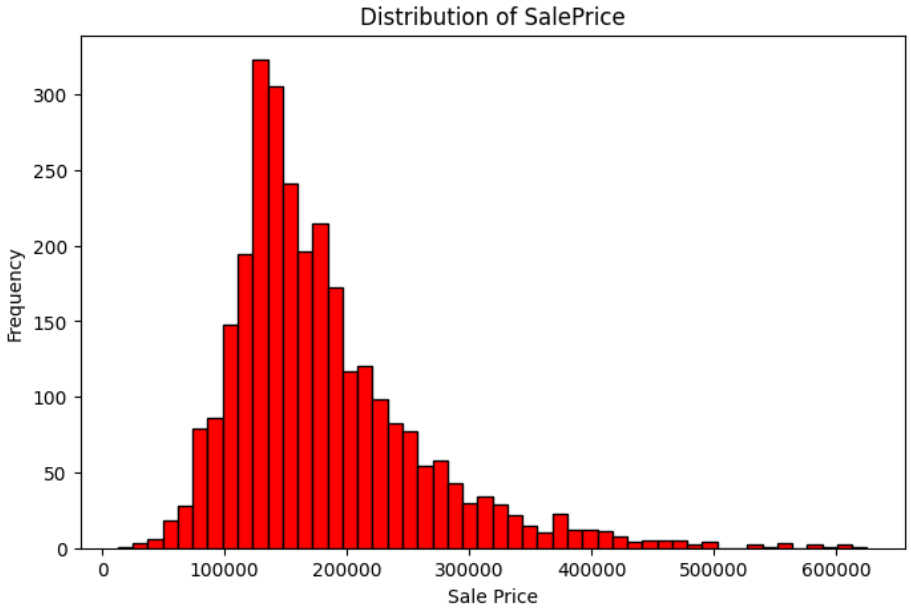
Each dummy equals 1 if the house lies in that neighborhood and 0 otherwise.

Note: The sum of Parts I–III is 21 + 1 + 25 = **47 features.**

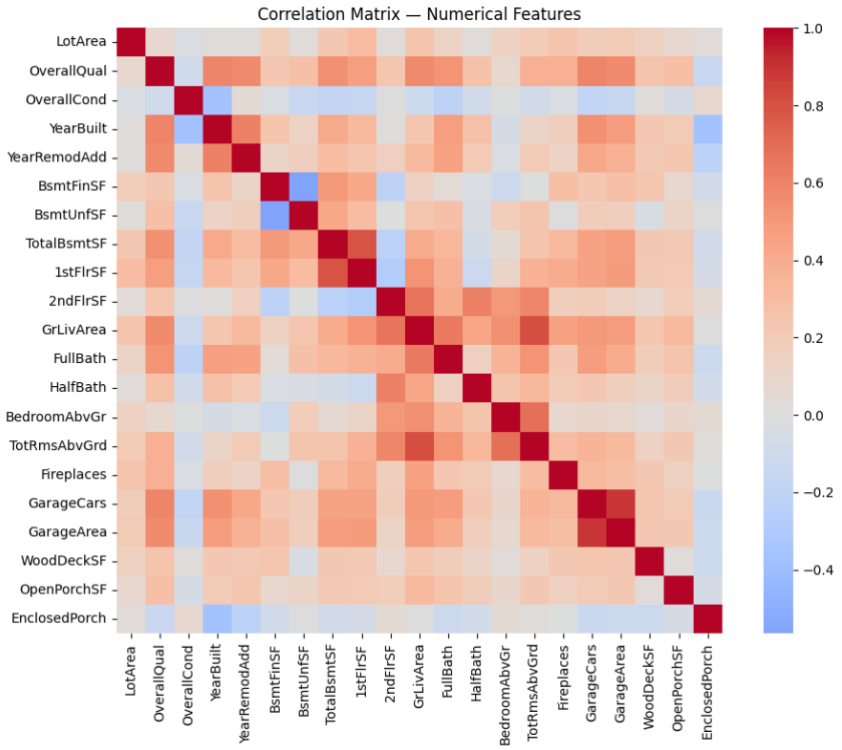
2.2 Data Cleaning

After preprocessing, no missing values remain in the modeling frame. In the raw data, the string "NA" in “BsmtQual” denotes “No Basement” rather than a missing observation; we explicitly

recode this level to 0 in the ordinal mapping described above. Visual and summary diagnostics did not reveal influential outliers requiring removal.



Multicollinearity is a consideration in linear models. Pairwise correlations among our selected predictors are modest, and model regularization (Ridge/Lasso) further mitigates collinearity concerns; therefore, we do not apply additional treatments.



2.3 Log-Transform of the Target

House prices are typically skewed and exhibit heteroskedasticity. To stabilize variance and reflect proportional effects, we model the log price:

$$y^* = log\ (y)$$

This transformation reduces the influence of high-end outliers and makes coefficients interpretable in approximate percentage terms. For reporting in dollar units, predictions are back transformed via

$$y = e^{y^*}$$

2.4 Normalization of Continuous Features

To place continuous variables on comparable scales, we standardize the 13 size-related measures:

{LotArea, YearBuilt, YearRemodAdd, BsmtFinSF, BsmtUnfSF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, GrLivArea, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch}

For each feature X_i, the standardized value is:

$$Z_i = \frac{X_i - \mu}{\sigma}$$

where μ is the mean and μ is the standard deviation. Categorical variables (the ordinal 'BsmtQual' code and the neighborhood dummies) are not normalized.

Note: To avoid data leakage, this will be applied after we split the original data set.

3. Model Building Method

We estimate three linear models: ordinary least squares (OLS) as a baseline, and its regularized variants—Ridge (l_2) and Lasso (l_1).

3.1 Data Split

The dataset is randomly partitioned into three disjoint sets: training ($n = 1800$), validation ($n = 600$), and test ($n = 508$). The training set is used to fit candidate models; the validation set is used for model selection, and the test set is held out strictly for the final performance assessment.

3.2 Evaluation Metrics

We compare models using mean squared error (MSE) and the coefficient of determination (R^2). Model selection is based on validation MSE (primary); R^2 is reported as a complementary goodness-of-fit measure.

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(Y_i - \tilde{Y}_i)^2$$

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

3.3 Regularization

For Ridge and Lasso, we conduct a hyperparameter search over predefined penalty values (λ). For each λ, we refit the model on the training set and evaluate on the validation set. The best model is the specification (method and λ) achieving the lowest validation MSE.

3.4 Final Assessment

After selecting the best specification on the validation set, we evaluate it once on the test set to report out-of-sample MSE and R^2.

4. Models

4.1 Linear Regression (OLS)

Given observations, the linear model is:

$$\tilde{y}^{(i)} = \sum_{j=0}^{n} \theta_j x_j^{(i)} = h_\theta\left(x_j^{(i)}\right), i = 1,2,3 \dots m$$

OLS estimates θ by minimizing the following cost function:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))^2$$

We fit OLS on the training set and report validation performance as follows:

Dataset	***MSE***	***RMSE***	$\boldsymbol{R^2}$
train	0.013628	0.116741	0.913317
val	0.015038	0.122628	0.911614

4.2 Ridge Regression (Ridge)

Ridge augments the OLS objective with a penalty to shrink coefficients and mitigate multicollinearity:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))^2 + \lambda \sum_{j=1}^{n} \theta_j^2$$

*Note: with the intercept excluding from penalization. (This scaling matches the implementation in **scikit-learn**, which uses $\frac{1}{2m}$ for the data-fit term.)*

We consider $\lambda \in \{0.10, 0.30, 0.60\}$.

We get the following results:

Lambda $(\boldsymbol{\lambda})$	***Dataset***	***MSE***	***RMSE***	$\boldsymbol{R^2}$
$\mathbf{0.10}$	$Train$	0.013628	0.116741	0.913317

	$Validation$	0.015038	0.122628	0.911614
$\mathbf{0.30}$	$Train$	0.013629	0.116742	0.913315
	$Validation$	0.015038	0.122629	0.911612
$\mathbf{0.60}$	$Train$	0.013630	0.116746	0.913309
	$Validation$	0.015039	0.122633	0.911607

4.3 Lasso Regression (Lasso)

This is similar to Ridge Regression, but it adds a penalty by using the absolute value of the coefficients. The cost function is given by:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))^2 + \lambda\sum_{j=1}^{n}|\theta_j|$$

We will apply $\lambda = 0.02,\ 0.06,\ 0.10$ respectively.

The results we get are as follows:

$\boldsymbol{Lambda\ (\lambda)}$	$\boldsymbol{Dataset}$	$\boldsymbol{MSE}$	$\boldsymbol{RMSE}$	$\boldsymbol{R^2}$
$\mathbf{0.02}$	$Train$	0.018560	0.136234	0.881952
	$Validation$	0.020906	0.144589	0.877121
$\mathbf{0.06}$	$Train$	0.029244	0.171008	0.813996
	$Validation$	0.032289	0.179691	0.810217
$\mathbf{0.10}$	$Train$	0.042960	0.207269	0.726752
	$Validation$	0.047086	0.216993	0.723245

4.4 Model Selection (Validation)

All candidate models (OLS, Ridge with the three λ's, and Lasso with the three λ's) are trained in the training split. We compute validation mean squared error (MSE) and select the specification (method and λ) with the lowest validation MSE. R^2 is reported as a complementary goodness-of-fit metric.

#	***Model***	***λ (Lambda)***	***Valid MSE***	***Valid RMSE***	***Valid R^2***
1	*Ridge*	0.10	0.015038	0.122628	0.911614
2	*Ridge*	0.30	0.015038	0.122629	0.911612
3	*Ridge*	0.60	0.015039	0.122633	0.911607
4	*Lasso*	0.02	0.020906	0.144589	0.877121
5	*Lasso*	0.06	0.032289	0.179691	0.810217
6	*Lasso*	0.10	0.047086	0.216993	0.723245

By comparison, we noticed that the **Ridge Regression with** $\lambda = 0.10$ provides us with the best model.

4.5 Final Test Evaluation

The selected model is evaluated once on the held-out test split. For our application with a log-price target, the test results are:

Metric	***Value***
MSE	0.015903
RMSE	0.126106

MAE	0.087897
R^2	0.905188

1. **R^2 (R-Squared): 0.905:**
   - **Interpretation:** This is the most telling metric, which is also known as the "coefficient of determination". It explains how much of the variance in the target variable is explained by our model.
   - **Scale:** 0 to 1. Closer to 1 is better.
   - **Our Score (0.905):** Our model explains 90.5% of the variance in the test data. This is an excellent result.

2. **RMSE (Root Mean Squared Error): 0.126**
   - **Interpretation:** This is the standard deviation of the prediction errors (residuals). It tells us the typical magnitude of the error our model makes, in the same units as our target variable. It penalizes large errors more heavily.
   - **Our Score (0.126):** On average, our model's predictions are about 0.126 units away from the actual values. This is a low value, indicating high precision.

3. **MSE (Mean Squared Error): 0.159**
   - **Interpretation:** This is the average of the squares of the errors. It is primarily used for model optimization (e.g., in the loss function).
   - Our Score (0.0159): The very low value confirms the excellent performance.

4. **MAE (Mean Absolute Error): 0.088**
   - Interpretation: This is the average of the absolute differences between the predicted and actual values. It gives us a more direct sense of the average error size.
   - Our Score (0.088): On average, our model's predictions are off by 0.088 units. This is very low and reinforces the conclusion from the RMSE.

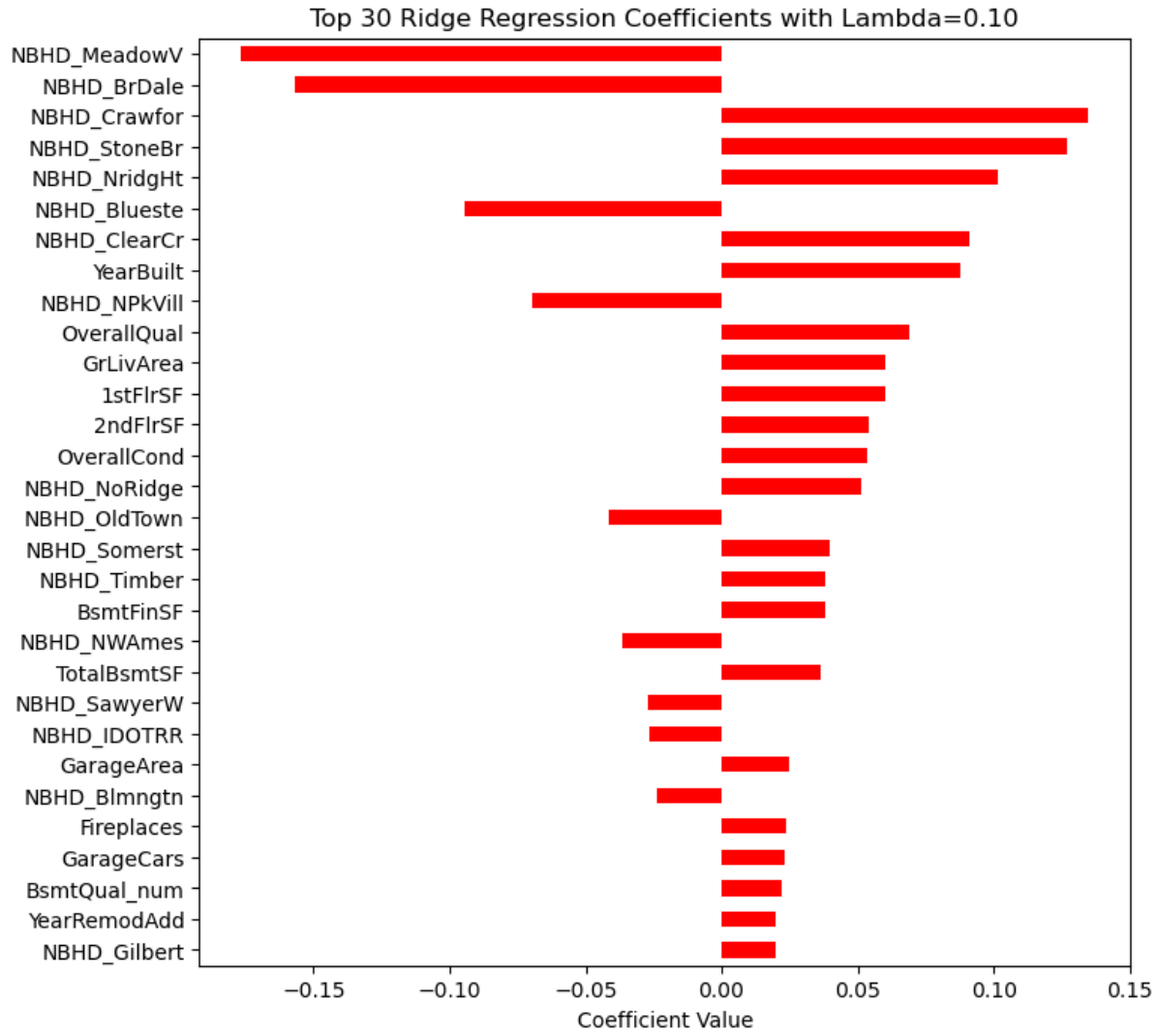
To summarize, we have built a highly accurate and robust predictive model. The results are consistent across all standard regression metrics, which gives high confidence in its performance. If we were predicting house prices, our model's predictions would, on average, be within 8.8% (MAE) of the actual sale price, and it captures over 90% of the factors that drive price fluctuations.

5. Findings

Now we will explain our key findings, focusing on which features contribute most to our model. And how they influence house prices. To have a clear view, we visualize the important features and their coefficients. The coefficients are shown below.

Feature	*Weights for ridge regression* ($\lambda = 0.10$)
Lot area (square feet)	0.019
Overall quality (scale: 1 to 10)	0.069
Overall condition (scale: 1 to 10)	0.054
Year built	0.088
Year remodeled (= year built if no remodeling or additions)	0.020
Finished basement (square feet)	0.037
Unfinished basement (square feet)	-0.004
Total basement (square feet)	0.036
First floor (square feet)	0.060
Second floor (square feet)	0.054
Living area (square feet)	0.060
Number of full bathrooms	0.004
Number of half bathrooms	0.014
Number of bedrooms	-0.000
Total rooms above grade	0.000
Number of fireplaces	0.024
Parking spaces in garage	0.023
Garage area (square feet)	0.025
Wood deck (square feet)	0.005
Open porch (square feet)	0.011
Enclosed porch (square feet)	0.012
Neighborhood (25 features)	-0.176 to 0.135
Basement quality	0.022

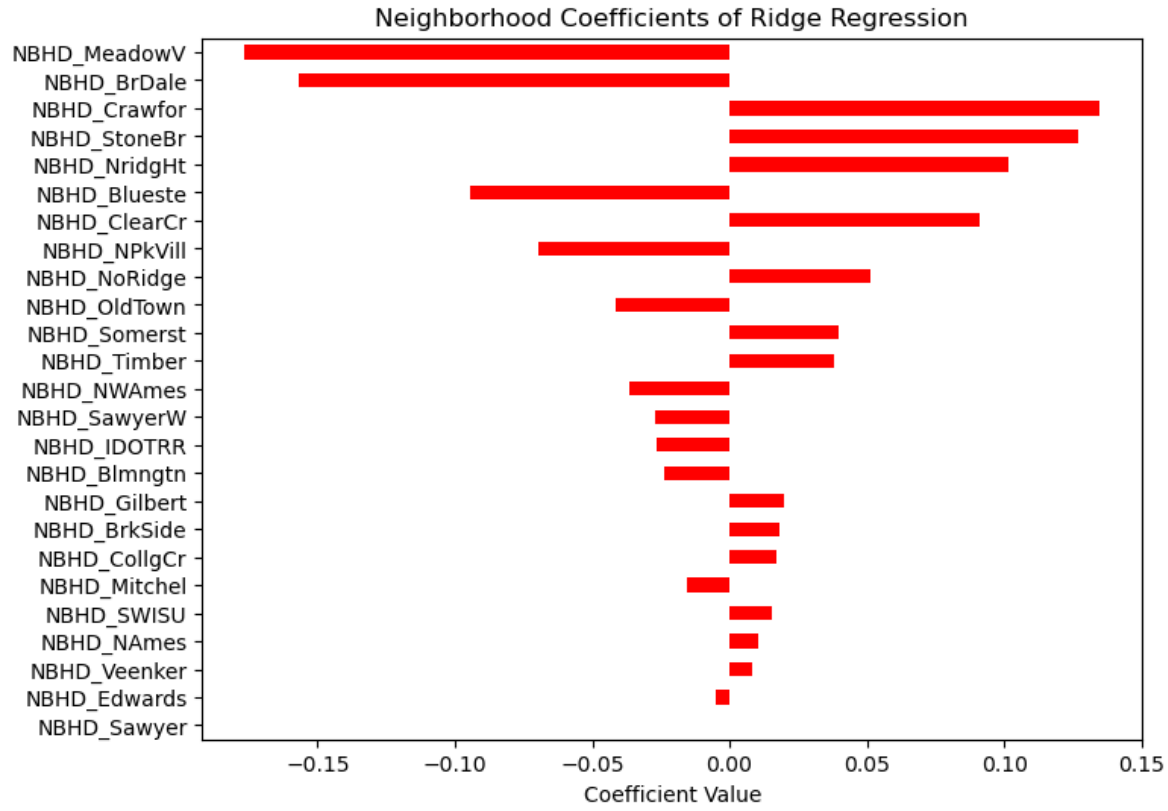
5.1 Top Features



Key Findings:

- **Location dominates.** *Neighborhood* indicators have the largest coefficients, indicating that location is the strongest predictor of log-price. This is consistent with industry intuition - “location, location, location.”
- **Quality & Size:** Beyond location, *Overall Quality* and interior living area (*1st/2nd floor*) exhibit large positive effects, as do *Finished Basement* and *Garage* measures. Interpreted on the log scale, these positive coefficients imply that higher quality and greater usable space are associated with higher prices, all else equal. Again, this aligns with our analysis.

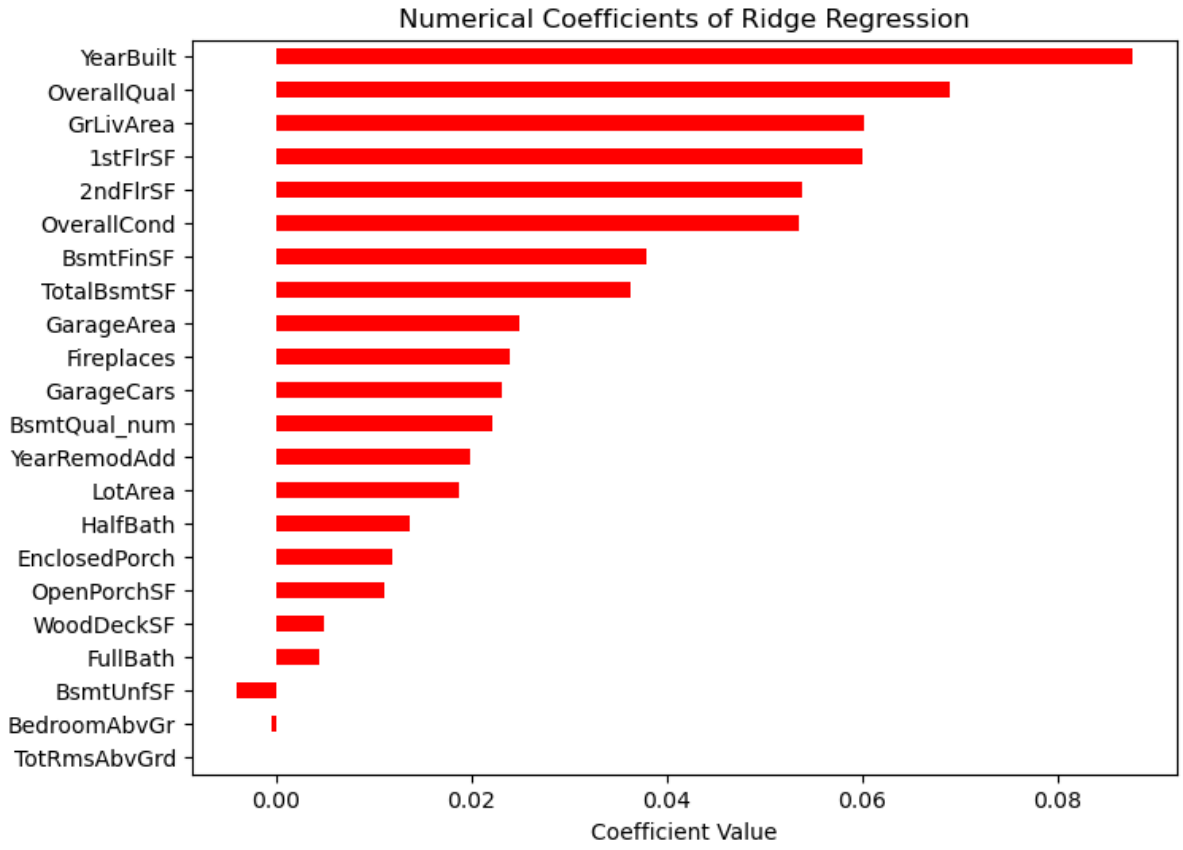
5.2 Neighborhood



Key Findings:

- **Some neighborhoods have positive effects on house prices while others have negative effects:** some areas are associated with price premiums, while others exhibit discounts. This pattern likely reflects differences in amenities and disamenities, e.g., infrastructure quality, public services, school catchments, and proximity to employment or recreation. In our estimates, the two largest negative effects occur in Meadow Village (MeadowV) and Briardale (BrDale). The strongest positive effects are in Crawford (Crawfor), Stone Brook (StoneBr), and Northridge Heights (NridgHt).

5.3 Numerical Features



Key Findings:

- **Original construction data matters:** YearBuilt carries a sizable positive coefficient, indicating that newer homes command higher prices, consistent with better building standards, systems, and finishes. Recent construction typically implies lower deferred maintenance and updated utilities, which buyers capitalize into price.
- **Quality and interior size dominate:** OverallQual is among the largest positive effects, followed by GrLivArea, 1stFlrSF, and 2ndFlrSF. On the log-price scale, these coefficients imply that higher quality and greater usable living area translate into meaningful percentage premiums, holding other attributes constant.
- **Basement and garage space add meaningful value; lot size less so:** TotalBsmtSF, BsmtFinSF, GarageArea, and GarageCars are clearly positive, reflecting premiums for finished/usable space and parking capacity. By contrast, LotArea is only modestly positive once interior area and neighborhood are controlled, suggesting buyers value interior livability more than raw lot size.
- **Negative effects:** BsmtUnfSF is slightly negative, which is economically intuitive: unfinished square footage adds less functional value and can even displace finished space, yielding a small discount relative to otherwise similar homes.

6. Predicting House Prices

We now deploy the selected model to generate real-time house price predictions using publicly available listings data from Zillow.com. For each property, we extract the required predictors (e.g., interior living area, basement and garage measures, quality indicators, and neighborhood identifiers), and compute the predicted log-price, which is then back-transformed to dollars.

$$y = e^{y^*}$$

Below is the link to this house.

https://www.zillow.com/homedetails/316-Topaz-Ct-Ames-IA-50010/93957355_zpid/

The key information about this house listed below:

Feature	***Value***	***Note***
LotArea	9583.2	*Sq ft*
OverallQual	9	*Estimated by the pictures online*
OverallCond	9	*Estimated by the pictures online*
YearBuilt	2003	
YearRemodAdd	2003	
BsmtFinSF	1434	*Sq ft*
BsmtUnfSF	0	*Sq ft*
TotalBsmtSF	1434	*Sq ft*
1stFlrSF	1000	*Estimated based on the design of the house*
2ndFlrSF	1491	*Estimated based on the design of the house*
GrLivArea	2491	
FullBath	2	
HalfBath	2	
BedroomAbvGr	4	
TotRmsAbvGrd	15	
Fireplaces	1	*We only found 1 fireplace*
GarageCars	2	*Estimated based on the picture of the garage*

$GarageArea$	600	$Sq\ ft$
$WoodDeckSF$	0	$We\ did\ not\ find\ any\ wood\ deck$
$OpenPorchSF$	200	$Estimated\ based\ on\ the\ picture$
$EnclosedPorch$	0	$We\ did\ not\ find\ any\ enclosed\ porch$
$BsmtQual\_num$	5	$Estimated\ based\ on\ the\ quality\ of\ the\ whole\ proper$

We used the model to calculate the predicted target value, and the formula to back-transform to dollars. The result we get is: $428,943.

And the price listed on Zillow is $468,000. Our prediction error can be calculated using the formula below, which is 8.35%.

$$\frac{|Actual\ Price - Predicted\ Price|}{Actual\ Price}$$

This means that our model performed well in predicting Iowa house prices.

7. Reference

1. John Hull, *Machine Learning in Business: An Introduction to the World of Data Science*, 3rd ed. (2021), 64–69.

2. Gopinath Rebala, Ajay Ravi, and Sanjay Churiwala, *An Introduction to Machine Learning* (Cham: Springer, 2019), 49–53.