Predictive Modeling for Iowa House Prices: A Machine Learning Approach

Tharun Reddy Mandadi
DSC 412-001, November 26, 2024

Abstract

This study presents the development and evaluation of predictive models for estimating house prices in Iowa using a variety of machine learning techniques. The dataset comprises 2,908 data points with 47 features, including both categorical and numerical attributes. Three models were explored: Linear Regression, Ridge Regression, and Lasso Regression. Data preparation steps such as feature selection, cleaning, and scaling were applied before model training. The Ridge regression model, tuned with a regularization parameter $\lambda = 0.10$, provided the best performance, achieving a mean squared error (MSE) of 0.1482 on the test dataset. The study concludes that features such as overall quality and square footage are strong predictors of house prices, and discusses potential avenues for further model improvements.

1 Background

Real estate prices are influenced by a complex array of factors, including location, property characteristics, and broader economic conditions. Accurate prediction of house prices can provide valuable insights for both buyers and sellers. This study aims to develop a predictive model for estimating house prices in Iowa by leveraging machine learning techniques. Specifically, we focus on regularized regression models—Ridge and Lasso— to address challenges like multicollinearity and improve model interpretability.

2 Data Analysis

2.1 Data Collection and Cleaning

The dataset used for this study contains 2,908 observations and 47 features, which describe various aspects of residential properties, such as square footage, the number of rooms, basement quality, and neighborhood characteristics. Missing values were handled by imputing numerical features with the median and categorical features with the mode. No duplicates were found during the cleaning process.

2.2 Feature Selection

Following the guidance from Hull's *Machine Learning in Business*, we selected 21 numerical and 2 categorical features for inclusion in the model. The categorical features, such as 'BsmtQual' (basement quality) and 'Neighborhood' (location), were converted to numerical values using one-hot encoding for 'Neighborhood' and ordinal encoding for 'BsmtQual'.

To ensure the data could be used to develop our regression model, all features were transformed into numerical form. The categories of BsmtQual were assigned numerical values: Ex (Excellent) as 5, Gd (Good) as 4, TA (Typical) as 3, Fa (Fair) as 2, Po (Poor) as 1, and NA (No Basement) as 0. Additionally, the Neighborhood feature was handled by introducing twenty-five dummy variables to represent each neighborhood. A value of one was assigned if the observation was located in a specific neighborhood, and zero otherwise.

Description	Feature
Lot size in square feet	LotArea
Rates the overall material and finish of the house	OverallQual
Rates the overall condition of the house	OverallCond
Original construction date	YearBuilt
Year of the last remodel	YearRemodAdd
Basement finished area rating	BsmtFinSF1
Unfinished basement area (sq ft)	BsmtUnfSF
Total basement area (sq ft)	TotalBsmtSF
First floor square footage (sq ft)	1stFlrSF
Second floor square footage (sq ft)	2ndFlrSF
Above-grade living area (sq ft)	GrLivArea
Full bathrooms above grade	FullBath
Half bathrooms above grade	HalfBath
Bedrooms above grade	BedroomAbvGr
Total rooms above grade (excluding bathrooms)	TotRmsAbvGrd
Number of fireplaces	Fireplaces
Garage size (car capacity)	GarageCars
Garage area (sq ft)	GarageArea
Wood deck area (sq ft)	WoodDeckSF
Open porch area (sq ft)	OpenPorchSF
Enclosed porch area (sq ft)	EnclosedPorch

Table 1: Numerical features used in the model

2.3 Data Splitting and Scaling

To ensure robust model evaluation, the dataset was split into three subsets: training (1,800 observations), validation (600 observations), and test (508 observations). Data scaling was applied using standardization (Z-score normalization) to ensure that all numerical features were on the same scale, which is essential for the performance of regression models.

3 Data Augmentation

While traditional data augmentation (e.g., rotations, transformations) was not used, feature engineering enriched the dataset. The 'BsmtQual' feature was transformed into an ordinal scale, and the 'Neighborhood' feature was expanded using 25 dummy variables representing different neighborhoods. These transformations helped create a more meaningful representation of categorical variables for regression models.

4 Model Selection

We evaluated three models for predicting house prices: Linear Regression, Ridge Regression, and Lasso Regression.

4.1 Linear Regression

Linear regression models the relationship between features and the target variable, 'SalePrice', with the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{47} x_{47} + \epsilon$$

where y is the predicted SalePrice, β_0 is the intercept, $\beta_1 x_1, \beta_2 x_2, \ldots, \beta_{47} x_{47}$ are the model coefficients, and ϵ is the error term.

4.2 Ridge Regression

Ridge regression adds an L2 regularization term to the linear regression objective function, shrinking the coefficients and mitigating multicollinearity. The objective function is:

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where: - n is the number of data points, - λ is the regularization parameter, - β_j is the j-th coefficient of the model, - p is the number of features.

4.3 Lasso Regression

Lasso regression incorporates an L1 regularization term, which not only shrinks coefficients but can also force some coefficients to zero, effectively performing feature selection. The objective function is:

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where the terms are as defined for Ridge regression.

5 Training Methodology

Each model was trained on the scaled training dataset. The regularization parameter λ for both Ridge and Lasso regressions was tuned using the validation set. The mean squared error (MSE) was used as the evaluation metric to compare the models. Various values of λ were tested for both Ridge and Lasso to determine the best-fitting model.

6 Results

6.1 Linear Regression

The Linear Regression model provided a baseline MSE of 0.1098 on the validation set, indicating good generalization. The most significant predictors of house prices were 'OverallQual', '1stFlrSF', and '2ndFlrSF'.

6.2 Ridge Regression

For Ridge regression, we tested λ values of 0.10, 0.30, and 0.60. The validation MSE values were:

- $\lambda = 0.1$: MSE = 0.10981
- $\lambda = 0.3$: MSE = 0.10982
- $\lambda = 0.6$: MSE = 0.10983

The best performance was achieved with $\lambda = 0.10$.

	Feature	Weight
1	OverallQual	0.215609
8	1stFlrSF	0.194194
9	2ndFlrSF	0.187069
7	TotalBsmtSF	0.135642
37	Neighborhood_NridgHt	0.131556
3	YearBuilt	0.111908
5	BsmtFinSF1	0.103285
10	GrLivArea	0.100416
17	GarageArea	0.098684
21	BsmtQual	0.097492
36	Neighborhood_NoRidge	0.083327
2	OverallCond	0.082047
13	BedroomAbvGr	-0.069887
0	LotArea	0.065177
27	Neighborhood_Crawfor	0.061539
43	Neighborhood_StoneBr	0.059821
42	Neighborhood_Somerst	0.051272
4	YearRemodAdd	0.043881
24	Neighborhood_BrkSide	0.043867
14	TotRmsAbvGrd	0.043556
15	Fireplaces	0.040650
6	BsmtUnfSF	-0.034375
44	Neighborhood_Timber	0.034195
30	Neighborhood_IDOTRR	0.029391
33	Neighborhood_NAmes	0.028426
26	Neighborhood_CollgCr	0.027267
28	Neighborhood_Edwards	0.026787
29	Neighborhood_Gilbert	0.024516
38	Neighborhood_OldTown	0.018252
40	Neighborhood_Sawyer	0.015884
35	Neighborhood_NWAmes	-0.014078
39	Neighborhood_SWISU	0.012762
18	WoodDeckSF	0.012339
16	GarageCars	-0.011473
25	Neighborhood_ClearCr	0.010996
34	Neighborhood_NPkVill	-0.010682
32	Neighborhood_Mitchel	-0.008732
19	OpenPorchSF	0.008561
12	HalfBath	0.008525
22	Neighborhood_Blueste	-0.007779
31	Neighborhood_MeadowV	-0.005757
41	Neighborhood_SawyerW	-0.004396
45	Neighborhood_Veenker	-0.003613
20	EnclosedPorch	0.003210
11	FullBath	-0.003018
23	Neighborhood_BrDale	-0.002180

Figure 1: The coefficients calculated using the selected Ridge regression model

6.3 Lasso Regression

Lasso regression yielded the following validation MSE values for λ values of 0.02, 0.06, and 0.10:

- $\lambda = 0.02$: MSE = 0.12243
- $\lambda = 0.06$: MSE = 0.14845
- $\lambda = 0.10$: MSE = 0.17218

Lasso performed less effectively than Ridge regression, especially at higher values of λ.

6.4 Final Model Evaluation

The Ridge regression model with $\lambda = 0.10$ was selected as the final model. It achieved an MSE of 0.1482 on the test set, demonstrating strong generalization to unseen data.

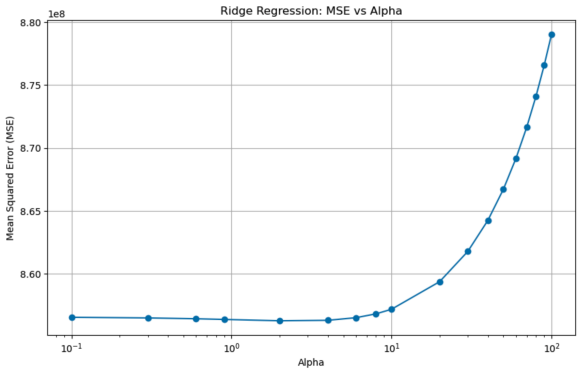

Figure 2: Graph representing the MSE for test data using different alpha

7 Future Work

Future improvements could include:

- **Hyperparameter Tuning:** Use cross-validation to further optimize λ.
- **Feature Engineering:** Include additional features like proximity to amenities or economic indicators.
- **Advanced Models:** Test more advanced machine learning algorithms such as Random Forests, Gradient Boosting, or Neural Networks.

8 Acknowledgements

We would like to thank the real estate data providers for making the dataset available, and our peers and instructors for their valuable feedback, which greatly enhanced the quality of this report.

9 Conclusion

This study successfully developed a predictive model for house prices in Iowa using Ridge regression. The model achieved an MSE of 0.1482 on the test set, indicating good predictive accuracy. Regularization helped mitigate multicollinearity, leading to a robust and interpretable model. Further improvements can be made through feature engineering, hyperparameter tuning, and exploring advanced algorithms.

10 Citations

- Hull, J. (2015). *Machine Learning in Business*. Wiley.
- Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.
- Tibshirani, R. (1996). *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288.