Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 8
AICT 2024

Feature Selection and Regression for House Value Prediction

Zhaowen Gu

Boston University, Boston, 02215, USA

Abstract. This report investigates the application of feature selection and regression techniques, including LASSO, Ridge Regression, and Elastic Net, in predicting house values. By analyzing these methods, we aim to identify the most influential factors driving house prices and assess their relative importance. Our study involves a comprehensive evaluation of various linear regression models to determine their effectiveness and consistency in predicting house prices. The results reveal that while all examined techniques demonstrate robust performance, certain methods, particularly Elastic Net, offer superior predictive accuracy and better handle multicollinearity among features. This report provides valuable insights for real estate professionals and data scientists seeking to refine their predictive models and make informed decisions based on comprehensive data analysis.

Keywords: Feature selection, Regression, House value prediction.

1. Introduction

Accurate prediction of house prices is a crucial aspect of the real estate industry, influencing decisions for buyers, sellers, and investors. The valuation of properties is influenced by a multitude of factors including location, size, age, amenities, and market conditions. As the real estate market becomes increasingly data-driven, employing sophisticated analytical techniques to predict house prices has become essential for making informed decisions.

Feature selection and regression analysis are pivotal components of this predictive process. Feature selection involves identifying the most relevant variables that contribute to predicting house prices, which helps in reducing the dimensionality of the data and improving the efficiency and accuracy of predictive models. By focusing on the most significant features, analysts can build models that not only perform better but also provide more interpretable results.

In this report, we explore three prominent regression techniques: LASSO (Least Absolute Shrinkage and Selection Operator), Ridge Regression, and Elastic Net. Each of these methods offers unique strengths in handling different aspects of regression analysis:

LASSO: Known for its ability to perform feature selection by shrinking some coefficients to zero, thereby excluding non-essential features from the model. Ridge Regression: Addresses multicollinearity by applying L2 regularization, which penalizes large coefficients and helps stabilize the model. Elastic Net: Combines the penalties of LASSO and Ridge Regression, offering a balance between feature selection and regularization, and providing a flexible approach to regression modeling. Our investigation aims to assess the performance of these regression techniques in the context of predicting house values. By applying these methods to the California Housing dataset, we evaluate their effectiveness in identifying key factors influencing property prices and compare their predictive accuracy.

This report provides an in-depth analysis of the methodologies employed, the performance metrics used to evaluate the models, and the implications of our findings. We aim to offer insights into how different regression techniques can be leveraged to enhance predictive accuracy and contribute to more informed decision-making in real estate.

2. Literature Review

2.1 Foundational Studies

Regression Shrinkage and Selection via the Lasso (Tibshirani, 1996): This seminal work introduced LASSO (Least Absolute Shrinkage and Selection Operator) as a method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical models.

Ridge Regression: Biased Estimation for Nonorthogonal Problems (Hoerl & Kennard, 1970): Early work on Ridge Regression introduced a technique to address multicollinearity in regression models by imposing a penalty on the size of the coefficients.

2.2 Hybrid and Enhanced Methods

Regularization and Variable Selection Via the Elastic Net (Zou & Hastie, 2005): This method combines penalties from both LASSO and Ridge Regression to handle models with highly correlated predictors effectively. It overcomes some of the limitations of LASSO when dealing with data where the number of predictors exceeds the number of observations.

2.3 Comparative Analyses

"Best Subset, Forward Stepwise, or Lasso? Analysis and Recommendations Based on Extensive Comparisons" (Hastie, Tibshirani, & Tibshirani): This study provides a direct comparison between best subset selection, forward stepwise selection, and LASSO, offering insights into the trade-offs between computational efficiency, model accuracy, and ease of interpretation.

In summary, the comparison between feature selection and regularization methods is crucial to understand which methods are most effective under different circumstances, especially in terms of handling overfitting, interpretability, and dealing with high-dimensional data. This project on comparative analysis will help in choosing the right model based on specific data characteristics and desired outcomes, reinforcing the connection between theoretical research and practical application in predictive analytics.

3. Project Description

This project delves into the realm of feature selection and regression techniques applied to the task of predicting house values. Through the utilization of machine learning algorithms such as LASSO, Ridge Regression, and Elastic Net, alongside comprehensive feature selection strategies, we aim to uncover the most influential factors driving house prices. By leveraging these methods, we endeavor to create models that not only yield accurate predictions but also provide insights into the underlying dynamics of the real estate market.

4. Methodology

4.1 Elastic Net Loss function

$$w = \text{argmin}_w \left(\frac{1}{2n} \lVert y - Xw \rVert^2_2 + \l1\_norm + \l2\_norm \right)$$

The l1_norm is defined as:

$$\l1\_norm = \alpha\footnote{The regularization parameter.} \cdot \l1\_ratio\footnote{The ratio of L1 penalty in the regularization.} \cdot \lVert w \rVert_1\footnote{The L1\_norm of w}$$

The l2_norm is defined as:

$$\l2\_norm = 0.5 \cdot \alpha \cdot (1 - l1\_ratio) \cdot \lVert w \rVert^2_2\footnote{The squared L2\_norm of w}$$

4.2 LASSO(Least Absolute Shrinkage and Selection Operator) Loss function

$$w = \text{argmin}_w \left(\frac{1}{2n} \| y - Xw \|_2^2 + \alpha \| w \|_1 \right)$$

4.3 Ridge Regression Loss function

$$w = \text{argmin}_w \left(\frac{1}{2n} \| y - Xw \|_2^2 + 0.5 \alpha \| w \|_2^2 \right)$$

4.4 Sequential Feature Search Algorithm

Sequential Feature Search (SFS) is a widely used algorithm for feature selection in machine learning. It iteratively selects features to build a model, evaluating each feature's contribution to the model's performance. Below is the pseudocode for the Sequential Feature Search algorithm:

Initialize: Set of selected features $S = \{\}$, maximum number of features k Select the feature f not in S that maximizes some criterion (e.g., classification accuracy, regression error reduction) Add f to S

In this algorithm, the set S starts empty, and features are incrementally added until the desired number of features k is reached. At each iteration, the algorithm selects the feature that maximizes a chosen criterion and adds it to S. This process continues until S contains k features. This algorithm has a time complexity of $O(n^2)$

After performing sequential feature search algorithm, we are plainly going to train our model using ordinary least squares with selected features.

4.5 Exhaustive Feature Search Algorithm

Exhaustive Feature Search (EFS) is a brute-force approach to feature selection where all possible feature subsets are evaluated to find the optimal subset. Below is the pseudocode for the Exhaustive Feature Search algorithm:

Initialize: Set of selected features $S = \{\}$, maximum number of features k, best feature subset $S_{best} = \{\}$, best criterion value $c_{best} = -\infty$ Compute the criterion value c for subset S' Update S_{best} to S' Update c_{best} to c

In this algorithm, all possible subsets of features up to size k are evaluated, and the subset with the highest criterion value is selected as the best subset. This algorithm has a time complexity of $O(2^n)$

Exhaustive Feature Search guarantees finding the optimal subset within the search space, but it becomes computationally expensive as the number of features increases.

After performing exhaustive feature search algorithm, we are plainly going to train our model using ordinary least squares with selected features.

5. Performance Metrics

To rigorously assess the performance of our predictive models, we employ several key regression metrics that provide insights into various aspects of model accuracy and error:

Mean Squared Error (MSE): Mean Squared Error is calculated as the average of the squared differences between the actual values and the predicted values. It provides a measure of the overall prediction error of the model. MSE is particularly sensitive to outliers because the errors are squared,

which means that larger errors contribute disproportionately to the metric. This can be useful in scenarios where minimizing large errors is critical.

Mean Absolute Error (MAE): Mean Absolute Error is the average of the absolute differences between the actual values and the predicted values. Unlike MSE, MAE treats all errors equally without squaring them, which makes it a more interpretable measure of average error. MAE provides a straightforward interpretation of how much, on average, predictions deviate from actual values. It is less sensitive to outliers compared to MSE, which can be advantageous when dealing with datasets containing extreme values.

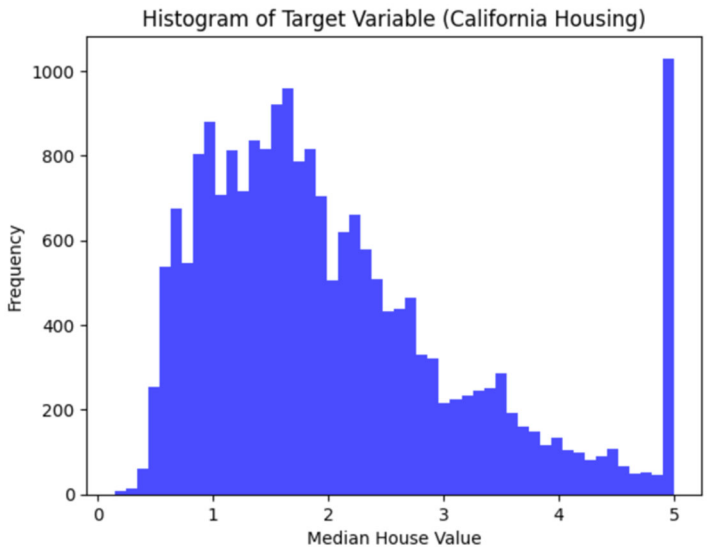
Coefficient of Determination (R^2): The R^2 score, also known as the coefficient of determination, measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. It ranges from 0 to 1, where a value of 1 indicates that the model explains all the variance in the target variable, and a value of 0 suggests that the model does not explain any of the variance. R^2 is a useful metric for understanding the goodness-of-fit of the model and its explanatory power.

Relative Absolute Error (R^1): Relative Absolute Error is calculated as the ratio of the sum of absolute errors to the sum of the absolute differences between the actual values and their mean. This metric provides insight into the model's performance relative to the scale of the target variable. It helps in understanding how well the model performs in the context of the range of the target values and is particularly useful when comparing models across datasets with different scales.

6. Dataset

We have utilized the California Housing dataset for our predictive modeling task. This dataset contains housing-related information for various districts in California, including features such as median income, housing median age, average rooms, average bedrooms, population, households, and geographical location. We need to predict the median housing price in hundreds of thousands.

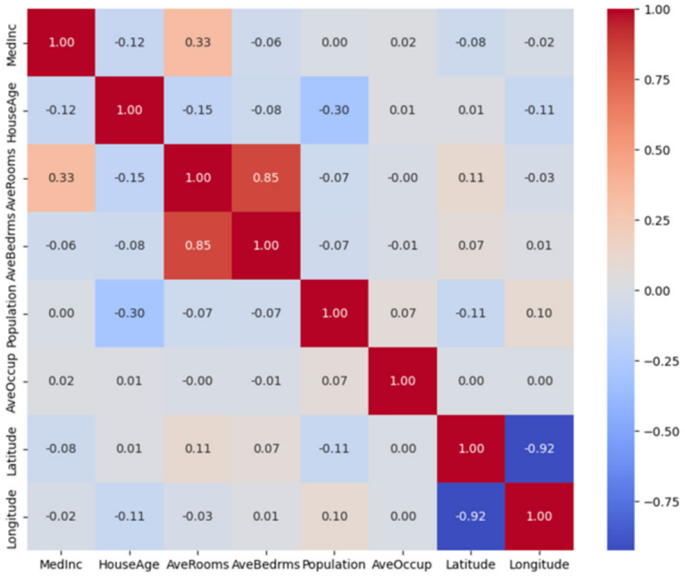
Histogram of target variable: median house value expressed in hundreds of thousands



Histogram of target variable: median house value expressed in hundreds of thousands

It is worth noticing that the target variable has a significant amount of outliers at 5 hundreds of thousands dollars.

Feature Correlation Matrix



Feature Correlation Matrix

We generated a feature correlation matrix to analyze the relationships between pairs of different features. In this matrix, red signifies strong positive correlations between feature pairs, while blue indicates strong negative correlations.

6.1 Data Preprocessing

To prepare the data for modeling, we performed the following preprocessing steps:

We normalized the feature matrix and target vector to ensure that all features are on a similar scale and have zero mean and unit variance. This step helps in stabilizing the training process and improving the convergence of the optimization algorithm.

We split the dataset into training and testing sets using a train-test split ratio of 80:20. This partitioning ensures that we have separate datasets for model training and evaluation.

The histograms of each normalized feature and the target variable y are included in the appendix. It's noteworthy that we inspected the histograms of the target variable y for both the training and testing portions to verify that they exhibit similar distributions. The histograms of the training and testing y values can also be found in the appendix. With the dataset preprocessed accordingly, it is now primed for training predictive models.

7. Experiments

7.1 Tuning hyperparameters

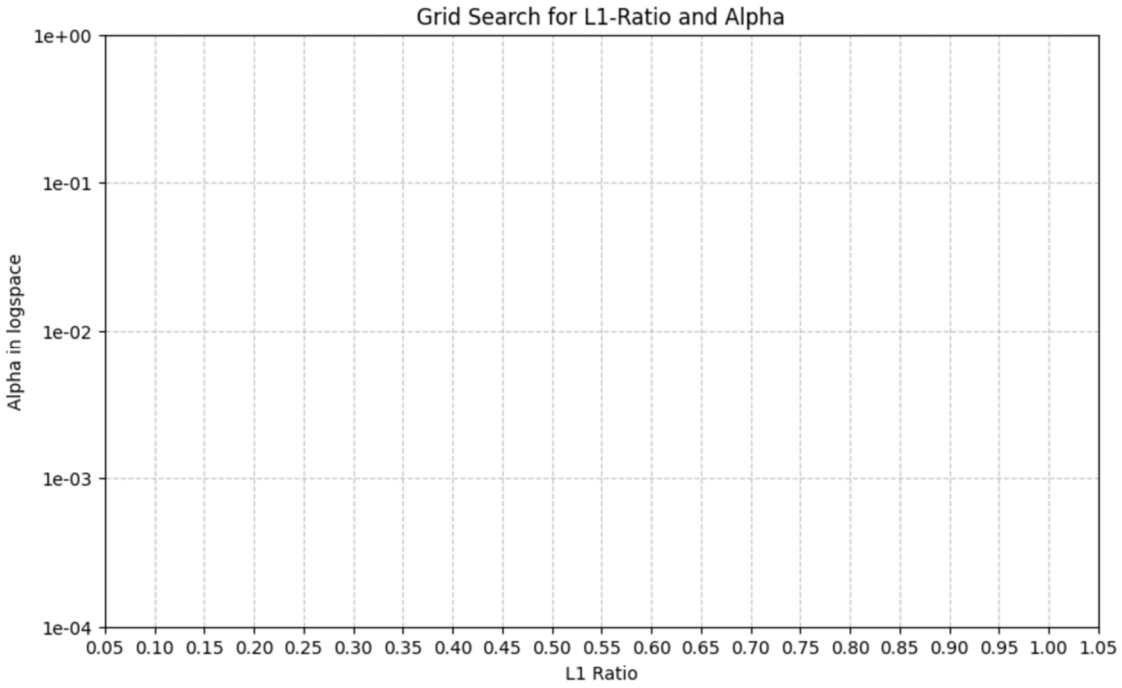
We are visualizing the performance metrics using a grid search for different values of hyperparameters L1_Ratio and Alpha. We are considering the following parameters:

$$\text{Alpha values: } [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0})$$

$$\text{L1 Ratio values: } [0.05, 0.1, 0.15, \ldots, 0.95)$$

We have initialized the Mean Squared Error (MSE) map with ∞ for each combination of Alpha and L1 Ratio. Then, we perform the grid search and update the MSE values accordingly.

Grid Search for L1_ratio and alpha



Grid Search for L1_ratio and alpha

In the visualization:

The x-axis represents the L1 Ratio values ranging from 0.05 to 0.95.

The y-axis represents the Alpha values in logspace from 10^{-4} to 10^{0}.

Each cell in the heatmap corresponds to the MSE value for a particular combination of Alpha and L1 Ratio.

During each iteration, we performed a 5-fold cross-validation to choose the best alpha and L1 ratio for Elastic Net regularization. The minimum MSE along with its corresponding alpha and L1 ratio values provide the best hyperparameters for the Elastic Net model.

Similarly, we are using the GridSearch with cross validation method to perform hyperparameter tuning for LASSO regression and Ridge regression regularization. There is only one hyperparameter to tune which is a 1-D vector alpha:

$$\text{Alpha values:} [10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1})$$

7.2 Training Process

We utilized various linear regression techniques for our predictive modeling task. Initially, we employed Sequential Feature Selection (SFS) and Exhaustive Feature Selector (EFS) in combination with a Ordinary Least Squares(OLS) model. Both SFS and EFS iteratively evaluated feature subsets to identify the optimal feature combinations. For each subset, the OLS model was trained and evaluated using 5-fold cross-validation, with the performance metric being the negative mean squared error (MSE).

Upon completion of the feature selection process, the OLS model was retrained using only the selected features. Subsequently, we employed Elastic Net, Lasso, and Ridge regression models individually without using the selected features. For Elastic Net, we utilized the previously determined optimal values of the regularization parameters (alpha and l1_ratio). Similarly, for Lasso and Ridge regression, we used the optimal alpha values obtained through grid search cross-validation.

We performed 5 fold cross-validation on these models and evaluated their performance using the same performance metrics as the OLS model (i.e. MSE). This allowed us to compare the performance of different regression techniques on the given dataset.

7.3 Testing

In the testing phase, we apply the testing dataset obtained from the earlier train-test split and generate predictions using each of the four linear models. Subsequently, we assess the accuracy of our predictions using the performance metrics outlined on page 3. These metrics include mean squared error, mean absolute error, coefficient of determination, and relative absolute error.

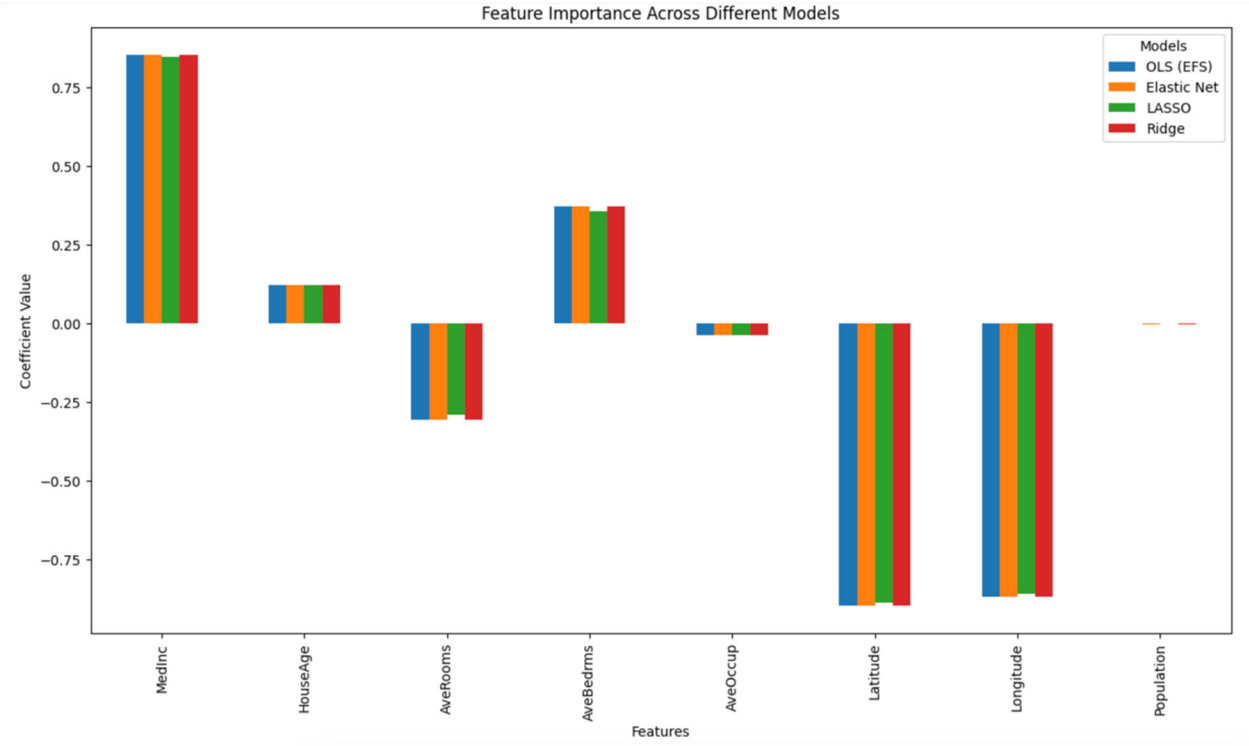
8. Results

For both SFS and EFS, the consensus was that all 7 features were selected except for the population of the block. Notably, SFS achieved this outcome in significantly less time compared to EFS, consistent with expectations.

Overall, our analysis yielded consistent performance across all four linear regression techniques:

These results indicate the robustness and effectiveness of linear regression techniques in predicting housing prices based on the provided features. However, further analysis and experimentation may be necessary to explore potential enhancements or alternative approaches for improving predictive accuracy.

Weight importance of each features



Weight importance of each features

Figure 4 captures the weight on each feature for all four linear regression models. Median income, latitude and longitude yields more feature importance while all four models put significantly less importance on population of the block and average number of house hold members.

9. Discussion

While our methods have demonstrated significant strengths in handling multicollinearity and enhancing model interpretability, our findings also suggest several areas for improvement.

9.1 Limitations of Current Approaches

Interpretation of Performance Metrics: The MAE for all models is 0.53, indicating an average gap of $53,000 between predicted and actual housing prices. This considerable disparity underscores the need for greater precision in our predictions. Additionally, an R-squared value of 0.58 implies that only 58% of the dataset's variability is accounted for by our model.

Small Number of Features: One of the critical limitations observed was the performance degradation when a small number of features were selected. While methods like LASSO inherently perform feature selection by reducing some coefficients to zero, this can sometimes lead to oversimplification, where potentially predictive features are excluded, thus limiting the model's ability to capture more complex patterns in the data.

Handling of Outliers: Another significant limitation was the sensitivity of linear methods to outliers in the dataset. Linear models, by their nature, can be significantly skewed by outliers, which affects their performance. Specifically, outliers stemming from peak housing pricing in the Bay Area highlighted the models' struggles with extreme values, which are common in real-world data.

9.2 Future Research Directions

Incorporation of Neural Networks: Considering the limitations of linear models, especially in handling non-linear relationships and outliers, adding neural networks to our comparison could provide deeper insights. Neural networks excel in capturing complex patterns and interactions in the data, making them suitable for datasets where traditional linear approaches underperform.

Comparative Analysis with Non-linear Methods: Extending our project to include a broader range of non-linear models, such as decision trees, random forests, and support vector machines, would provide a more comprehensive understanding of how different models perform across various datasets and feature sets.

10. Conclusion

In conclusion, our study delved into the realm of feature selection and regression techniques to predict house values. By leveraging machine learning algorithms like LASSO, Ridge Regression, and Elastic Net, along with comprehensive feature selection strategies, we aimed to uncover the most influential factors driving house prices.

Our analysis revealed that both Sequential Feature Selection (SFS) and Exhaustive Feature Selector (EFS) identified all features as significant except for the population of the block. Notably, SFS achieved this outcome more efficiently than EFS, as anticipated.

Across all four linear regression techniques, including Best Subset Selection with Ordinary Least Squares (OLS), Elastic Net, LASSO, and Ridge Regression, we observed consistent performance.

These findings underscore the robustness and effectiveness of linear regression techniques in predicting housing prices based on the provided features. However, our study also highlighted some limitations, such as the sensitivity of linear models to outliers and potential oversimplification when dealing with a small number of features.

Moving forward, incorporating neural networks and exploring a broader range of non-linear methods could provide deeper insights into predictive modeling for real estate valuation. Such approaches could offer enhanced capabilities in capturing complex patterns and handling outliers, ultimately leading to more accurate predictions.

References

[1] Regression Shrinkage and Selection via the Lasso (Tibshirani, 1996)

[2] Ridge Regression: Biased Estimation for Nonorthogonal Problems (Hoerl & Kennard, 1970)

[3] Regularization and Variable Selection Via the Elastic Net (Zou & Hastie, 2005)

[4] "Best Subset, Forward Stepwise, or Lasso? Analysis and Recommendations Based on Extensive Comparisons" (Hastie, Tibshirani, & Tibshirani)

[5] Gene selection for cancer classification using support vector machines (Guyon et al., 2002)

[6] Wrappers for feature subset selection (Kohavi & John, 1997)

[7] Regularization Paths for Generalized Linear Models via Coordinate Descent (Friedman et al., 2010)

[8] Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1), 55-67.

[9] Zou, H., & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.

[10] Hastie, T., Tibshirani, R., & Tibshirani, R. (2009). Best Subset, Forward Stepwise, or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *The Statisticians*, 2(3), 406-413.

[11] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, 46(1), 389-422.

[12] Kohavi, R., & John, G. H. (1997). Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97(1-2), 273-324.

[13] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1-22.

[14] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

[15] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273-297.

[16] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. *Springer*.

[17] Friedman, J., & Meulman, J. (2003). Multiple Additive Regression Trees with Applications to Heteroscedasticity. *Statistical Modelling*, 3(3), 191-210.

[18] Zhang, H., & Zhang, L. (2011). A Study of Feature Selection Based on Elastic Net Regularization. *Journal of Machine Learning Research*, 12, 3009-3036.

Appendix

Statistics of Features

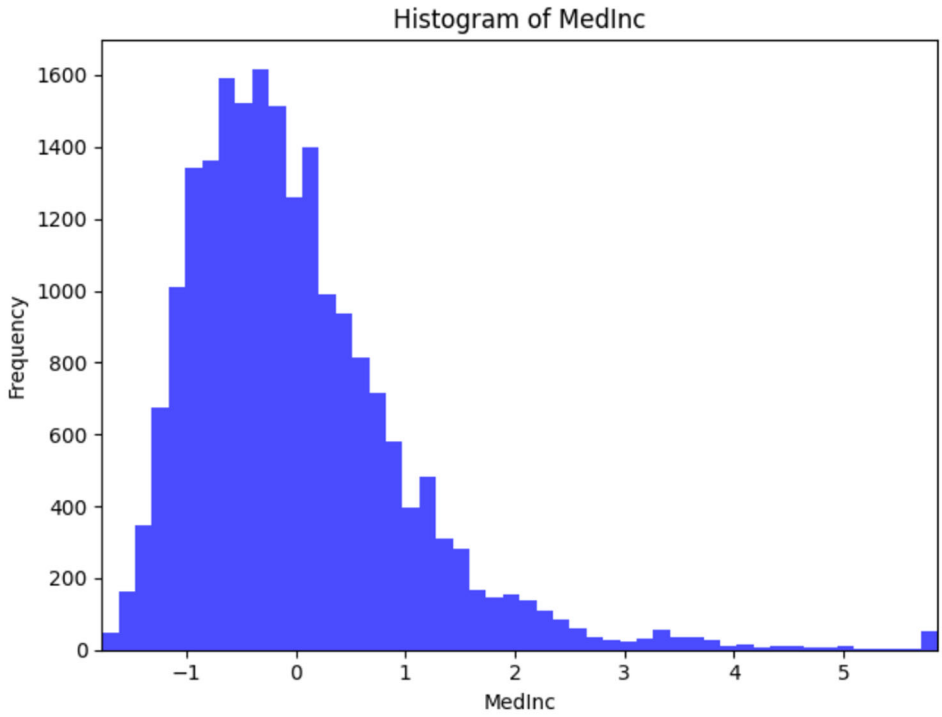
Statistics of Target Variable

Statistics of Target Variable

	Target
count	20640.00
mean	2.07
std	1.15
min	0.15
25%	1.20
50%	1.80
75%	2.65
max	5.00

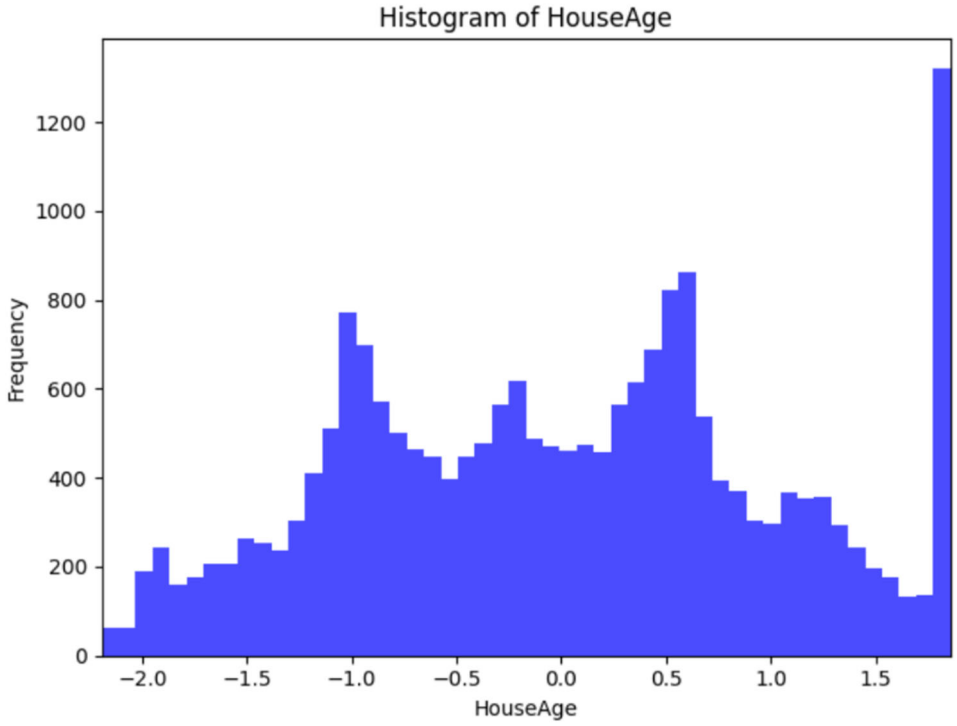
Histogram of normalized features

Histogram of normalized feature: median income in block group



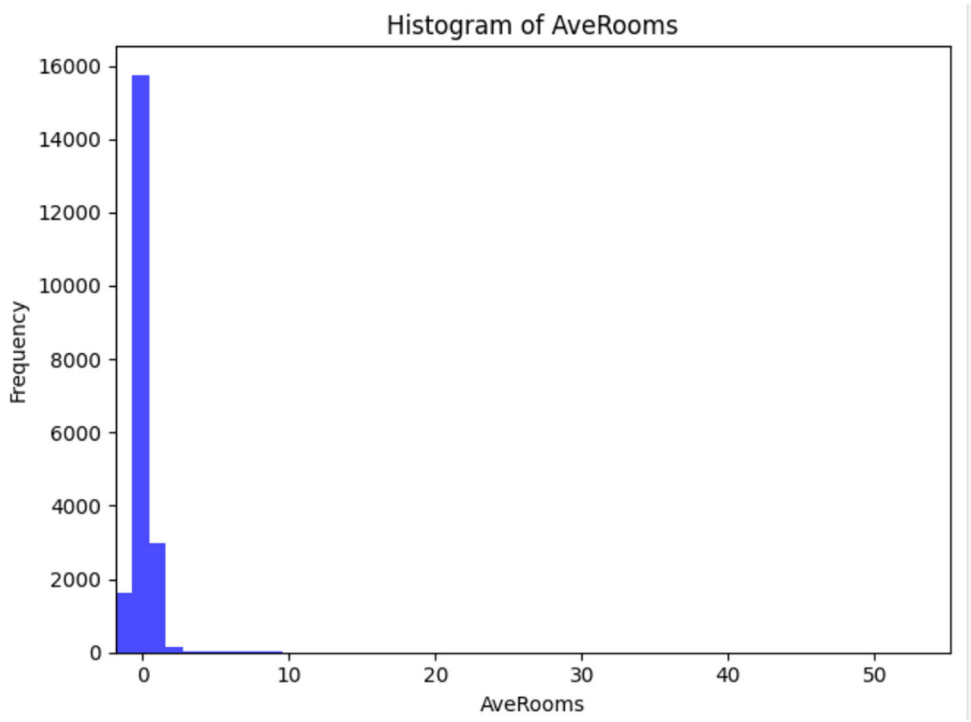
Histogram of normalized feature: median income in block group

Histogram of normalized feature: House age



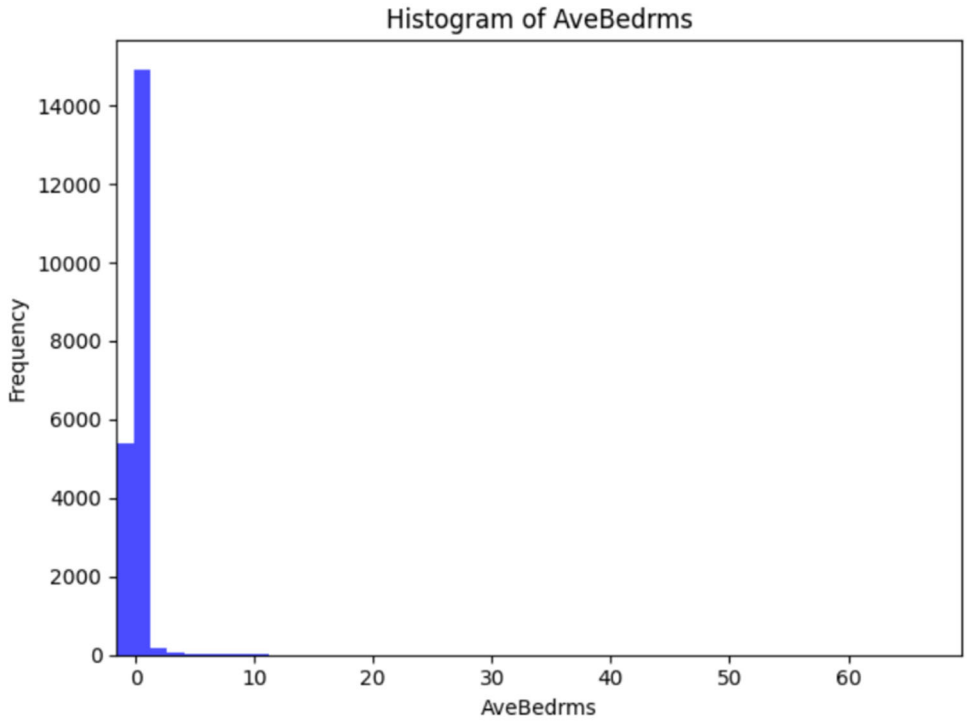
Histogram of normalized feature: House age

Histogram of normalized feature: Average number of rooms



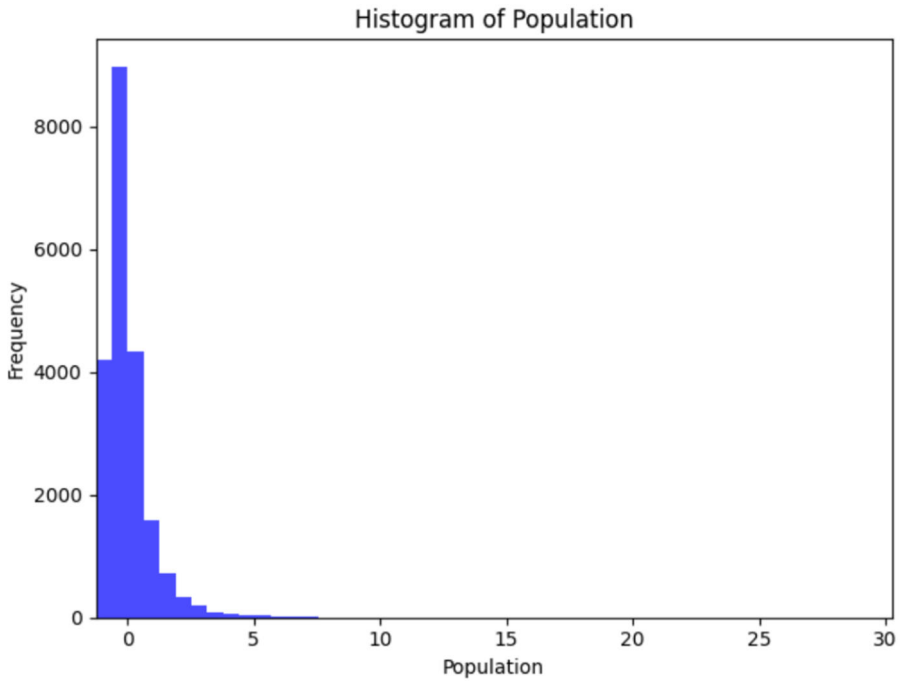
Histogram of normalized feature: Average number of rooms

Histogram of normalized feature: Average number of Bedrooms



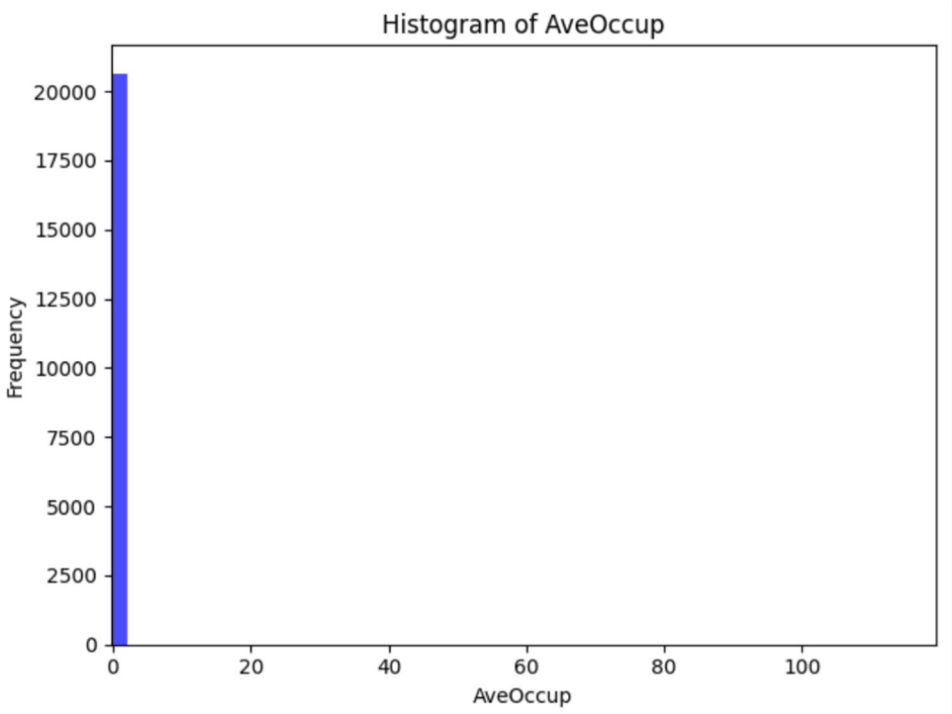
Histogram of normalized feature: Average number of Bedrooms

Histogram of normalized feature: Block group population



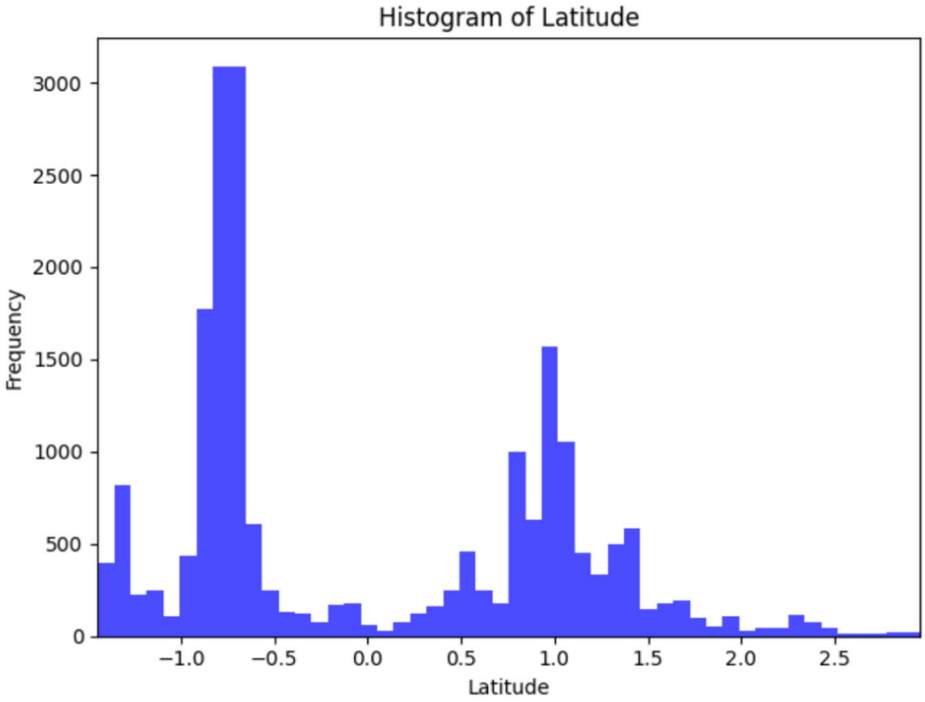
Histogram of normalized feature: Block group population

Histogram of normalized feature: Average number of household members



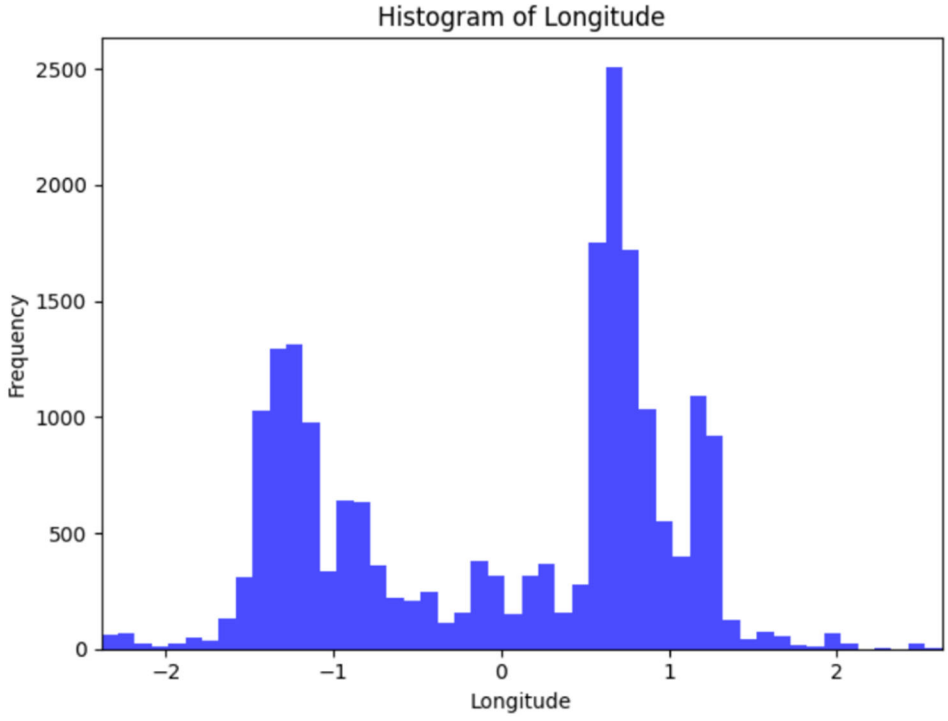
Histogram of normalized feature: Average number of household members

Histogram of normalized feature: Latitude



Histogram of normalized feature: Latitude

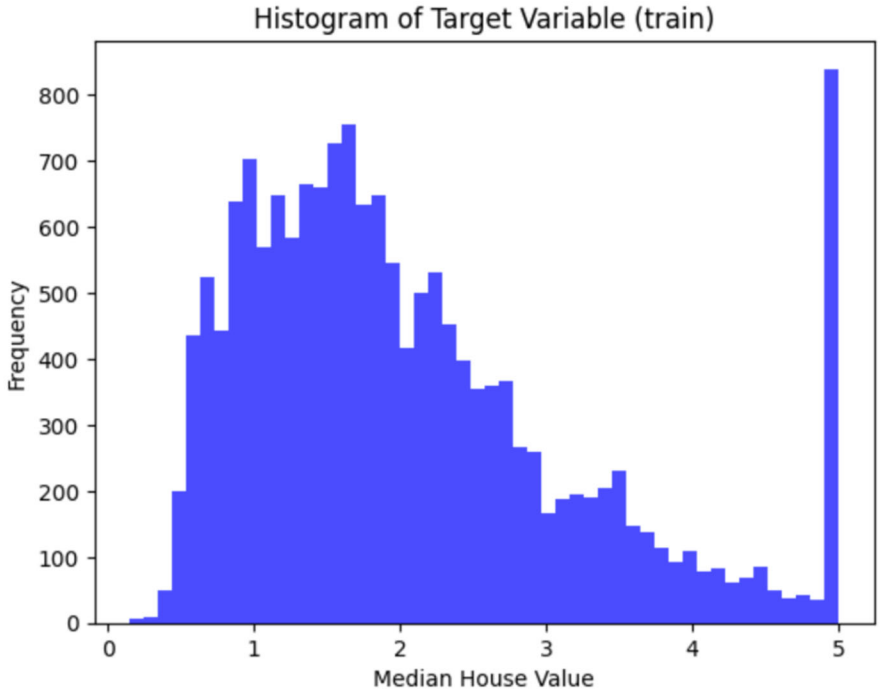
Histogram of normalized feature: Longitude



Histogram of normalized feature: Longitude

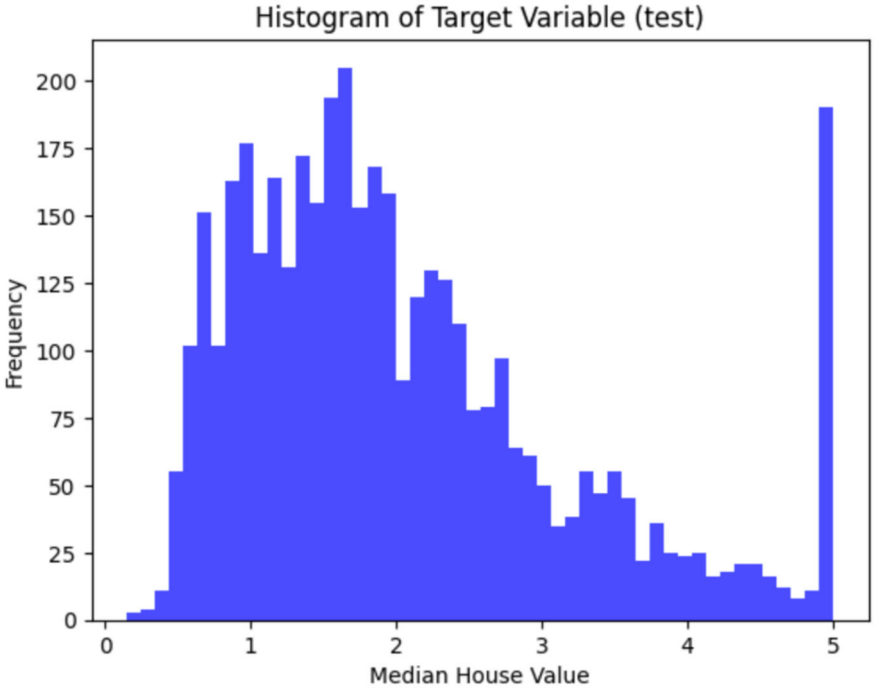
Histogram of training and testing target variable

Histogram of target variable in training dataset



Histogram of target variable in training dataset

Histogram of target variable in testing dataset



Histogram of target variable in testing dataset