A logo with black text

Description automatically generated

CS770 Machine Learning

Assignment 1: Linear Regression, Ridge & Lasso Regression

03/02/2024

Submitted by: Logan Schraeder (x356t577)

**Introduction**

Assignment 1 of the Spring 2025 semester focused on data preprocessing, exploratory data analysis (EDA), and simple and multivariate linear regression. Data preprocessing and EDA are absolutely vital steps at the beginning of any kind of data analysis or machine learning pipeline, even while regression or more advanced methods are the informing algorithms. The first steps of any data analysis or machine learning project involve data preprocessing and exploratory data analysis (EDA). These are essential for making sure the results of your analysis are reliable, even when using complex algorithms.

Data preprocessing involves cleaning and transforming raw data so it can be analyzed. This may include handling missing values, addressing outliers, and changing how data is displayed. EDA involves summarizing the data's main characteristics and finding potential patterns or relationships. This can include calculating summary statistics, visualizing distributions, and exploring correlations. By carefully performing these steps, one can understand the data, identify potential issues, and make informed decisions about which modeling techniques to use. This ultimately leads to extracting useful and actionable knowledge from the data. In Assignment 1, regression and its different techniques were the focus.

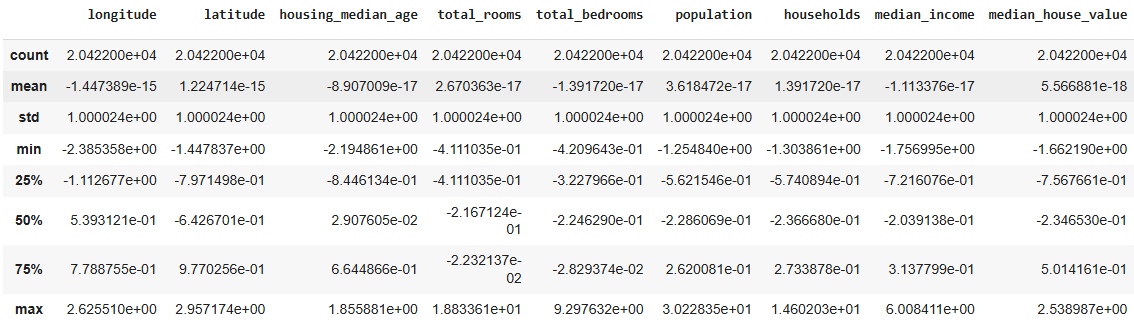
Once data preprocessing and EDA are complete, analysts can begin to build and train models. One of the simplest and most widely used modeling techniques is linear regression, which models the relationship between a dependent variable and one or more independent variables.

**Methods**

This assignment dictated the use of two overarching regression methods: simple and multivariate linear regression. Simple linear regression is a statistical method that models the relationship between a dependent variable and a single independent variable. It is used to predict the value of the dependent variable based on the value of the independent variable. Multivariate linear regression is an extension of simple linear regression that models the relationship between a dependent variable and two or more independent variables. It is used to predict the value of the dependent variable based on the values of the independent variables.

Within both regression types - but more widely used with multivariate regression - are two techniques called “Lasso” and “Ridge” regression. Lasso regression, also known as “Least Absolute Shrinkage and Selection Operator” regression, is a type of linear regression that uses a penalty term to shrink the coefficients of the independent variables (L1 regularization). This penalty term is proportional to the absolute value of the coefficients, which means that the coefficients are penalized more heavily for larger values. As a result, lasso regression tends to produce sparse models, with many of the coefficients being exactly zero.

Lasso regression is often used when there are a large number of independent variables and when some of the variables are correlated. This is because lasso regression can help to select the most important variables and to reduce the effects of collinearity. It is worth mentioning that in this assignment, normalization was used as a pre-processing step, further reducing the effectiveness of lasso.



**Figure 1. Post-Standardization Feature Descriptive Statistics for CA Housing.**

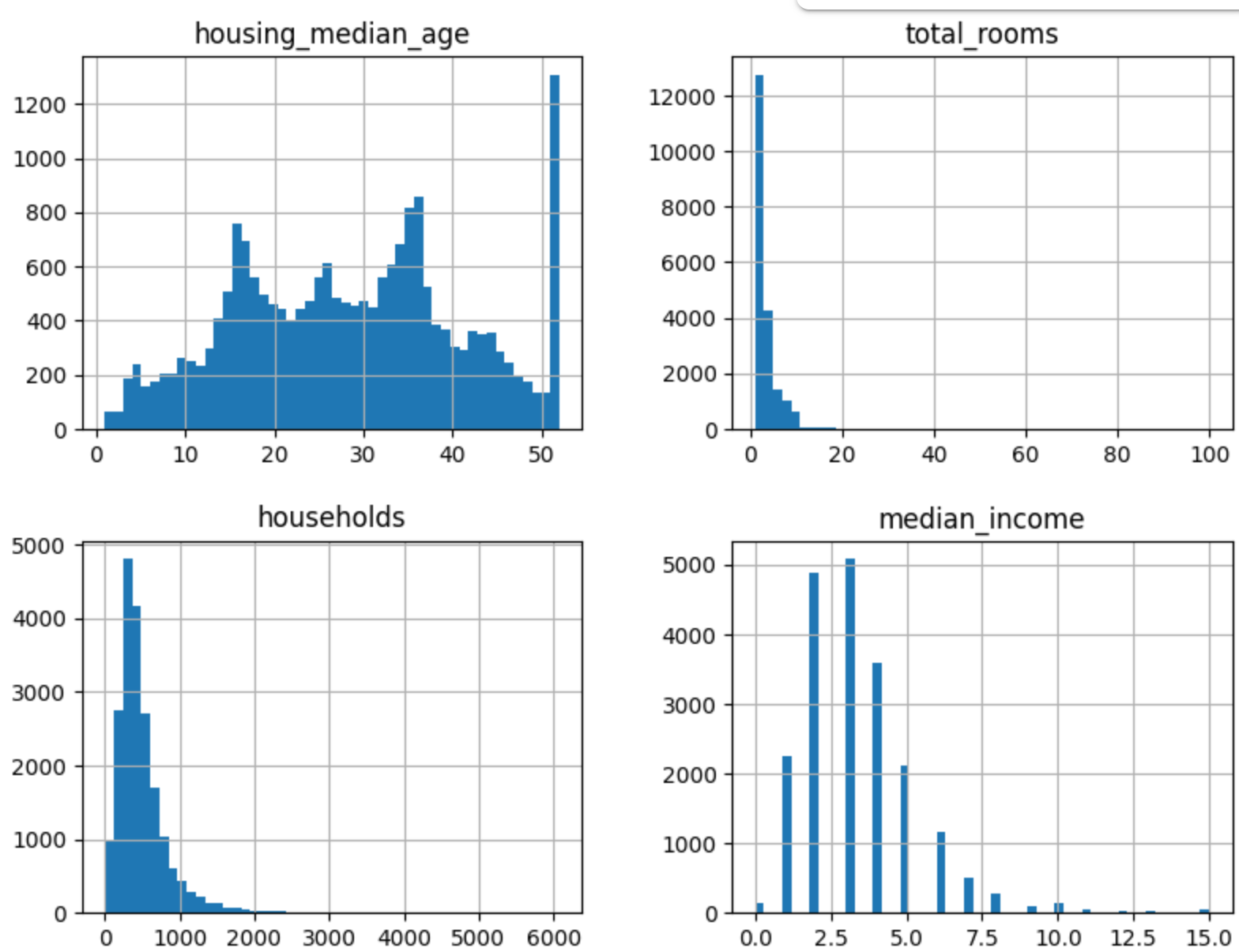
Ridge regression is a type of linear regression that uses a penalty term to shrink the coefficients of the independent variables (L2 regularization). This penalty term is proportional to the squared value of the coefficients, which means that the coefficients are penalized more heavily for larger values. As a result, ridge regression tends to produce models with smaller coefficients than ordinary least squares regression.

Ridge regression is often used when there is a high degree of collinearity among the independent variables. This is because ridge regression can help to reduce the effects of collinearity and to produce more stable models.

***Question 1 - CA Housing Dataset***

The California Housing Dataset EDA followed generally standard EDA steps. Many of the features displayed positively-skewed Gaussian distributions, while a handful were simple binary features. Two noteworthy events from the CA dataset EDA included:

1. The need for both the “Total Bedrooms” and “Total Rooms” features to be “scaled” even before standardization occurred. Most or all of the observations for these variables were unreasonable and would lead one to believe there was a data entry or format conversion error (e.g. 8000 total rooms instead of 8; 250 bathrooms instead of 2.5).
2. There were many features that displayed multicollinearity. When calculating the features’ Variance Inflation Factor (VIF), the most egregious were: latitude/longitude, and all ocean proximity variables.



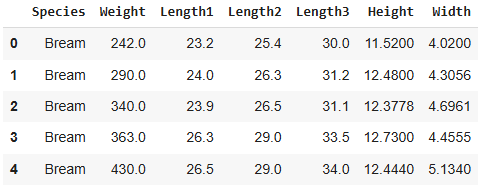
**Figure 2. Histograms and Distributions of Selected Features.**

The scaling of features was especially important here because of the presence of various features that exist in the real world on continuous-but-different scales. For example, the population of a municipality might be 150,000 people but the number of households in the same area is only 2200. They are both continuous values but are orders of magnitude different, and without proper scaling or standardization, they can negatively impact the performance of a model.

For model training and evaluation, the target variable was “Median House Value”, and the other features were used to attempt to predict it (the exception being simple linear regression, wherein only “population” was used to try and predict median house value). The two metrics used to evaluate performance of the models and feature selections was Mean Squared Error (MSE) and R2. While both metrics are fractional numbers, they represent very different things. R2 indicates how well a regression model actually fits the data (“goodness of fit”, with a value closer to 1 being better), while MSE is the average of the errors in predictions (with closer to 0 being better).

***Question 2 - Fish Dataset***

The same process as in Question 1 was followed for Question 2, the exception being that there did not appear to be any “scale data entry” errors. The same methods for regression were followed and will be discussed further in the “results” section to avoid redundancy.



**Figure 3. Fish Dataset Snapshot.**

**Results and Discussion**

***Question 1 - CA Housing Dataset***

The California Housing dataset was analyzed using exploratory data analysis (EDA) and linear regression modeling. Data preprocessing was performed to address data quality issues by correcting errors, handling missing values, and scaling features. Correlation and VIF were used to identify multicollinearity among independent variables. Outliers were detected and removed to improve the regression models.

Several linear regression models (OLS, Ridge, and Lasso) were trained to predict median house values; the single-variable standard linear regression model yielded a strong R² score of 0.001, indicating a very good fit. Regularized models, particularly Lasso, were very sensitive to the alpha parameter; lowering alpha values generally improved model performance, as shown by increased R² and decreased MSE (shown in Table 1).

**Question 2 - Fish Dataset**

The Fish dataset was analyzed using a linear regression model to predict fish weight based on length, height, and width. Data standardization was performed before modeling, and multivariate linear regression produced the highest R² value (0.882), indicating strong predictive ability. The model's coefficients show how each feature impacted fish weight. When Ridge and Lasso regularization were introduced, the results faired similar to those of the CA dataset; Lasso regression showed improved performance with reduced alpha values. These results demonstrate that linear regression, and its regularized variants, can effectively predict fish weight from dimensional measurements.

| **California Housing Dataset** | | | | **Fish Dataset** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Method** | **R^2** | **MSE** | **Alpha** | **Method** | **R^2** | **MSE** | **Alpha** |
| **Single** | 0.001 | 0.999 | N/A | **Single** | N/A | N/A | N/A |
| **Multi** | 0.634 | 0.368 | N/A | **Multi** | 0.882 | 0.132 | N/A |
| **Ridge** | 0.639 | 0.369 | 1.0 | **Ridge** | 0.877 | 0.138 | 1.0 |
| **Lasso** | 0.000 | 1.021 | 1.0 | **Lasso** | 0.549 | 0.504 | 0.5 |
|  | 0.208 | 0.809 | 0.5 |  | 0.753 | 0.276 | 0.3 |
|  | 0.426 | 0.586 | 0.3 |  | 0.838 | 0.181 | 0.15 |

**Table 1. Comparison of Regression Models.**

Across both datasets, linear regression proved to be a useful approach. Preprocessing and EDA were crucial in ensuring data quality and model reliability. Perhaps most interestingly - albeit intuitive - was that multivariate linear regression was a good tradeoff between training time and accuracy. Regularization techniques, especially Lasso, demonstrated the importance of parameter tuning in optimizing model performance.In theory, one could use code to iteratively loop through and find the optimal alpha value, but this was not done in the favor of time.

The strong R² values obtained across various models indicate that the chosen features effectively captured the relationships within the data, showing their ability to predict housing values and fish weight. An interesting and important note is the special sensitivity that Lasso regularization shows to learning rate, and an iterative approach to bounding alpha can prove particularly useful.

**Conclusions**

In conclusion, the analysis of both the California Housing and Fish datasets demonstrated the effectiveness of linear regression models and their different modalities in predicting target variables. The importance of disciplined preprocessing and exploratory data analysis (EDA) was quickly evident in the CA dataset and greatly impacted model reliability. Regularization techniques, especially Lasso regression, showed sensitivity to parameter tuning, with varying alpha values significantly impacting model performance. High R² values across multiple models indicate strong predictive capabilities. Future work could include optimizing learning parameters for Lasso and Ridge, optimizing feature selection, or introducing feature engineering.