Machine Learning Model Performance Report

1. Introduction

In this report, we explore the application of three regression techniques: Linear Regression, Lasso Regression (L1), and Ridge Regression (L2) on the California housing dataset. Each model is evaluated based on **Mean Squared Error (MSE)** and **R-Squared (R2)** metrics. The report highlights the improvements in model performance due to regularization, the impact of the regularization term (alpha), and provides a recommendation on the best-suited technique for this use case.

2. Methodology

The dataset was preprocessed by:

1. **Standardizing** numerical features to ensure consistency across features with varying scales.
2. Splitting the dataset into **training** (80%) and **testing** (20%) sets.

The following models were applied:

- **Linear Regression (LR)**: A standard regression model without regularization.
- **Lasso Regression (L1)**: A regression model that uses L1 regularization, which encourages sparsity in the model by driving some coefficients to zero.
- **Ridge Regression (L2)**: A regression model that uses L2 regularization, which shrinks the coefficients but does not eliminate any of them.

For **Lasso** and **Ridge**, hyperparameter tuning was performed using **GridSearchCV** to find the optimal alpha (regularization strength) values from a predefined set.

3. Results and Discussion

3.1 Model Performance

- **Mean Squared Error (MSE)**:
    - **Linear Regression**: 0.3578
    - **Lasso Regression**: 0.3653
    - **Ridge Regression**: 0.3578

Discussion: Both Linear Regression and Ridge Regression achieved similar MSE values, indicating that the models performed comparably. Lasso Regression showed a slightly higher MSE, which suggests that the regularization might have been too strong and caused underfitting.

- **R-Squared (R2)**:

- **Linear Regression**: 0.6346
- **Lasso Regression**: 0.6270
- **Ridge Regression**: 0.6346

Discussion: Both Linear Regression and Ridge Regression had the same R2 score, explaining around 63% of the variance in the data. Lasso performed slightly worse, suggesting that Lasso's feature selection mechanism might have excluded important features, thus affecting the model's ability to explain the variance.

3.2 Impact of Regularization (Alpha)

Regularization in both Lasso and Ridge aims to reduce overfitting by penalizing the model complexity. The key difference lies in how each regularizer handles the coefficients:

- **Lasso** (L1 regularization) can set some coefficients to zero, effectively performing **feature selection**. This can lead to sparse models, which might perform worse if important features are excluded.
- **Ridge** (L2 regularization) reduces the magnitude of the coefficients but does not set them to zero. It helps prevent large weights in the model, ensuring that the model is more robust but retaining all features.

The choice of alpha determines the strength of the regularization:

- A **higher alpha** value will result in stronger regularization, leading to a simpler model that may underfit if too many coefficients are shrunk or set to zero.
- A **lower alpha** value allows the model to retain more flexibility, which may lead to overfitting if regularization is too weak.

3.3 Comparison between Lasso and Ridge

- **Lasso Regression**:
    - Tends to exclude features with zero coefficients, which is useful for datasets with many irrelevant features.
    - However, it may perform worse in this case if important features are discarded during regularization.
- **Ridge Regression**:
    - Tends to perform better when all features are relevant, as it does not eliminate any features.
    - In this case, Ridge performed similarly to Linear Regression, suggesting that regularization provided slight improvements without overly penalizing the features.

4. Recommendation

Ridge Regression is recommended for this use case for the following reasons:

- **Stability**: Ridge provides stability by shrinking coefficients without eliminating features. It is ideal when all features are expected to contribute meaningfully to the model.
- **Performance**: Ridge and Linear Regression performed similarly, but Ridge offers better generalization by preventing overfitting, which might be especially important with a larger dataset or more features.
- **Feature Set**: Since no feature elimination was observed, Ridge is better suited for situations where all features are expected to play a role in predicting the target variable.

Lasso would only be recommended if feature selection is critical, or if you expect many irrelevant features in the dataset that need to be discarded.

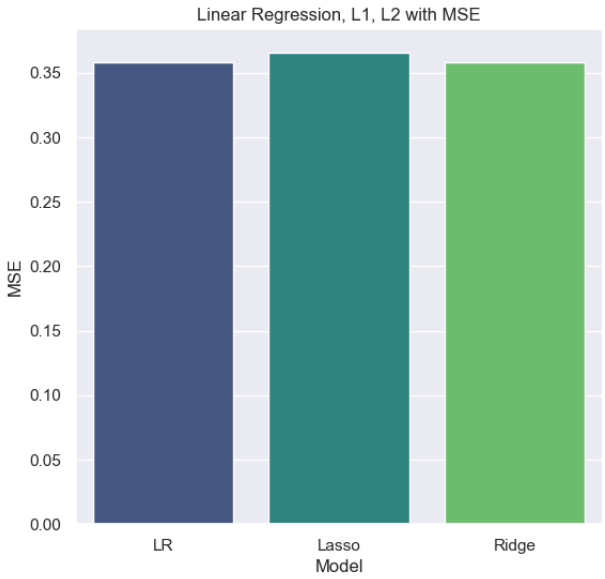
5. Conclusion

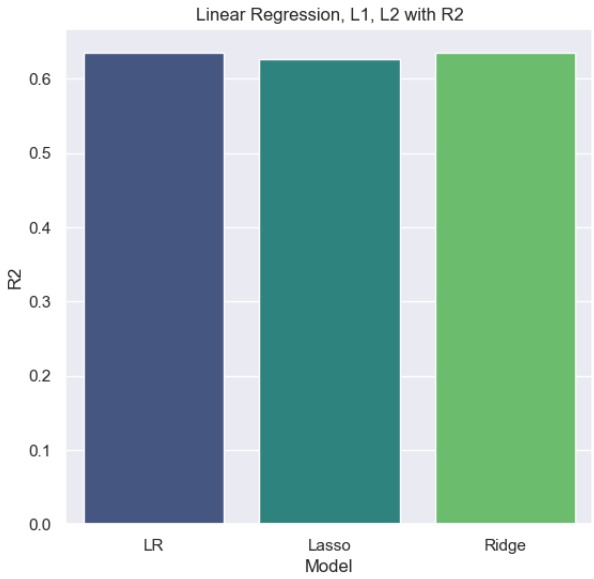
In conclusion, Ridge Regression offers the best balance between regularization and model performance for this dataset. The evaluation metrics suggest that Ridge is effective in handling potential overfitting, while Lasso's feature selection might lead to loss of relevant information.

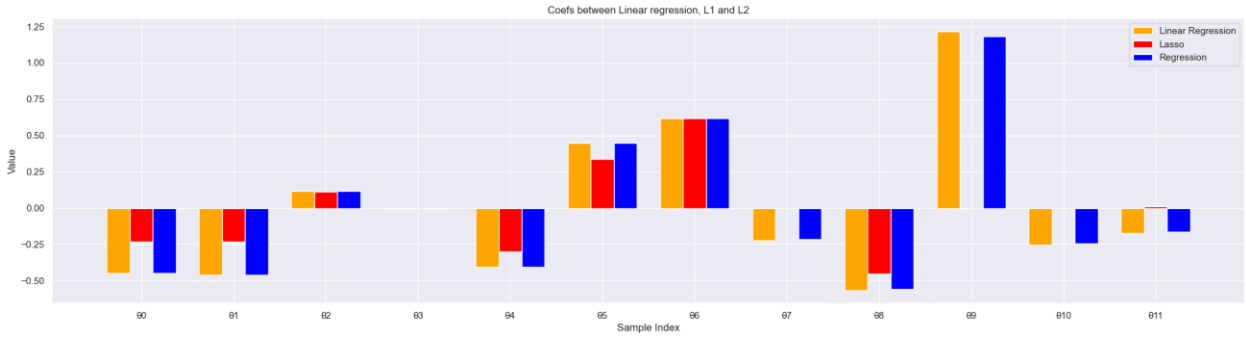
6. Visualizations

Figures:

1. **MSE Bar Plot**: Comparison of MSE values for Linear Regression, Lasso, and Ridge.
2. **R2 Bar Plot**: Comparison of R2 scores for the models.
3. **Coefficient Comparison**: Side-by-side comparison of coefficients for Linear Regression, Lasso, and Ridge to visualize the impact of regularization.


Linear Regression, L1, L2 with MSE
MSE
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
LR
Lasso
Ridge
Model


Linear Regression, L1, L2 with R2
R2
0.6
0.5
0.4
0.3
0.2
0.1
0.0
LR
Lasso
Ridge
Model


Coefs between Linear regression, L1 and L2
Linear Regression
Lasso
Regression
Value
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
θ0
θ1
θ2
θ3
θ4
θ5
θ6
θ7
θ8
θ9
θ10
θ11
Sample Index