Report on Housing Prices Prediction

1. Dataset and Features

The dataset used in this analysis is the Boston Housing dataset, which contains 506 observations of 13 features. The target variable is the median value of owner-occupied homes (MEDV). The features are as follows:

1. CRIM: Per capita crime rate by town.
2. ZN: Proportion of residential land zoned for lots over 25,000 sq. ft.
3. INDUS: Proportion of non-retail business acres per town.
4. CHAS: Charles River dummy variable (1 if tract bounds river; 0 otherwise).
5. NOX: Nitrogen oxides concentration (parts per 10 million).
6. RM: Average number of rooms per dwelling.
7. AGE: Proportion of owner-occupied units built prior to 1940.
8. DIS: Weighted distances to five Boston employment centers.
9. RAD: Index of accessibility to radial highways.
10. TAX: Full-value property tax rate per $10,000.
11. PTRATIO: Pupil-teacher ratio by town.
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of Black residents by town.
13. LSTAT: Percentage of lower status of the population.

2. Data Preprocessing Steps

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

1. Handling Missing Values: Checked for any missing values in the dataset. There were no missing values. However, there were a lot of missing values in the columns ZN and CHAS present as 0. Replaced those values with the median of the respective columns.

2. Feature Scaling: Standardized the features to ensure all variables contribute equally to the model. The Min max scaling was used for this purpose.

3. Train-Test Split: The dataset was split into training and testing sets with an 80-20 split ratio.

3. Model Training and Evaluation Results

Three regression models were trained and evaluated: Linear Regression, Ridge Regression, and Lasso Regression. The evaluation metrics used were Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared (R^2), and Adjusted R-squared (Adjusted R^2).

Linear Regression

- MAE: 0.0610
- MSE: 0.0072
- R^2: 0.6435
- Adjusted R^2: 0.5856

Ridge Regression
- Best Lambda: 1.0
- MAE: 0.0597
- MSE: 0.0067
- R^2: 0.6627
- Adjusted R^2: 0.6080

Lasso Regression
- Best Lambda: 0.0010
- MAE: 0.0636
- MSE: 0.0073
- R^2: 0.6325
- Adjusted R^2: 0.5730

4. Interpretation of the Model's Performance and Coefficients

Performance Interpretation

- **Ridge Regression** outperformed the other models, indicating it handles multicollinearity well by penalizing large coefficients, leading to a more stable and generalizable model.
- **Linear Regression** provided moderate performance but was susceptible to overfitting due to high variance.
- **Lasso Regression** had the weakest performance, possibly due to its strong penalty on coefficients, which led to zeroing out less significant features.

Coefficient Interpretation

Ridge Regression (Best Model):
- LSTAT and RM had the highest absolute coefficients, indicating they are the most significant predictors of housing prices.
- DIS, PTRATIO, and NOX also had notable coefficients, highlighting their influence on housing prices.
- Some features, like AGE and INDUS, had very small coefficients, suggesting a lesser impact on the target variable.

5. Challenges Faced

1. **Feature Selection**: Determining the most significant features was challenging due to multicollinearity.
2. **Hyperparameter Tuning**: Choosing optimal values for Ridge and Lasso regression hyperparameters required extensive cross-validation and computational resources.
3. **Model Interpretation**: Understanding the impact of each feature on the target variable, especially with penalized regression methods, was complex due to the interaction effects and penalties applied.

Conclusion

The Ridge Regression model provided the best predictive performance, with lower MAE and MSE and higher R^2 and Adjusted R^2 values compared to Linear and Lasso Regression models. The coefficients of the best model indicated that LSTAT, RM, DIS, PTRATIO, and NOX are significant predictors of housing prices