Report on ML Project for Housing Prices

1. Dataset and Its Features

Dataset: The dataset used in this analysis is the "Boston Housing Dataset," which contains information on housing prices in Boston suburbs. The dataset comprises 506 observations of 14 variables.as

Features: The dataset includes the following features:

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

2. Data Preprocessing Steps

Loading Data: The dataset was loaded into a pandas DataFrame, with appropriate column names assigned to each feature.

Handling Missing Values: No missing values were identified in the dataset, ensuring a complete dataset for analysis.

Outlier Detection: Outliers were detected using the Z-score method. A threshold of 3 was used, and rows with Z-scores exceeding this threshold were identified as outliers.

Feature Scaling: Two types of feature scaling were performed:

- **Standardization**: Using `StandardScaler` to scale the features to have a mean of 0 and a standard deviation of 1.
- **Normalization**: Using `MinMaxScaler` to scale the features to a range between 0 and 1.

Splitting Data: The data was split into training and testing sets using an 80-20 split.

3. Model Training and Evaluation Results

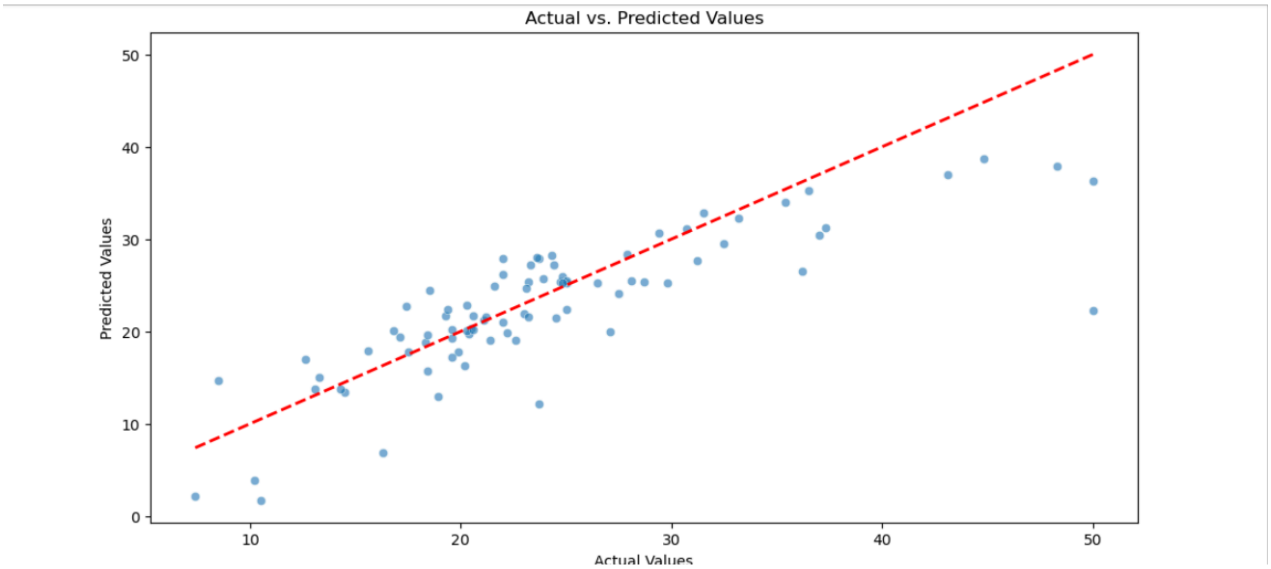
Model Training: Ridge Regression was used for model training. This method was chosen due to its ability to handle multicollinearity by adding a penalty to the size of the coefficients.

Evaluation Metrics:

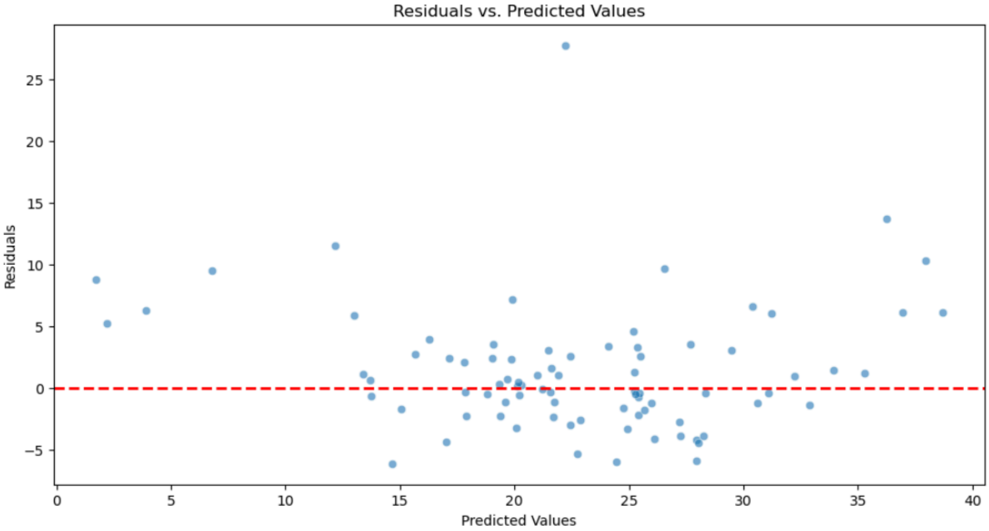
- **Mean Absolute Error (MAE)**: 3.235
- **Mean Squared Error (MSE)**: 20.198
- **R-squared (R2)**: 0.798

Visualization of Model Performance:

- **Scatter Plot of Actual vs. Predicted Values**: This plot indicated a strong positive correlation, with most points lying close to the 45-degree reference line.



- **Residual Plot**: The residuals were randomly scattered around zero, suggesting no obvious patterns and indicating a good model fit.

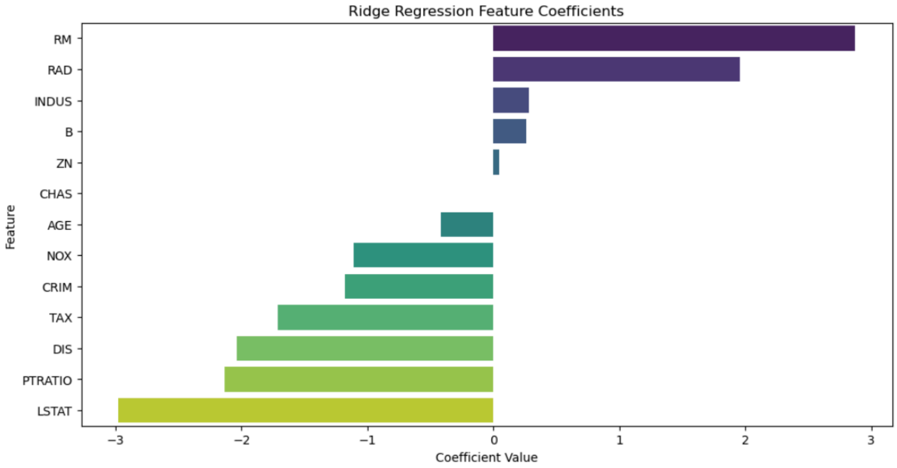


4. Interpretation of the Model's Performance and Coefficients

Model Performance: The R-squared value of 0.798 indicates that approximately 80% of the variance in housing prices can be explained by the model. The MAE and MSE values are relatively low, suggesting that the model's predictions are close to the actual values.

Coefficients: The coefficients from the Ridge Regression model provide insights into the impact of each feature on housing prices:

- **Positive Coefficients**: Features such as RM (average number of rooms per dwelling) and B (proportion of Black residents) had positive coefficients, indicating that increases in these features are associated with higher housing prices.
- **Negative Coefficients**: Features such as LSTAT (percentage of lower status population) and NOX (nitric oxide concentration) had negative coefficients, indicating that increases in these features are associated with lower housing prices.



Coefficient Shrinkage: Due to regularization, the coefficients are shrunk compared to a standard linear regression model. This reduces the impact of less important features and helps prevent overfitting.

5. Challenges Faced During the Task

Multicollinearity: Handling multicollinearity was a challenge, which was addressed by using Ridge Regression. This method helps by adding a penalty to the size of the coefficients, thereby reducing the impact of multicollinearity.

Outliers: Identifying and handling outliers required careful analysis. The Z-score method was used, but deciding on a threshold was crucial to avoid removing too many data points.

Feature Scaling: Choosing the right scaling method (standardization vs. normalization) required experimentation. Standardization was ultimately chosen for model training, but normalization was also considered.

Model Interpretation: Interpreting the coefficients in the presence of regularization required understanding how Ridge Regression affects the magnitude of the coefficients.

Conclusion

The Ridge Regression model provided a robust and interpretable analysis of the Boston Housing Dataset. By addressing multicollinearity and outliers, the model achieved good predictive performance and offered valuable insights into the factors influencing housing prices. Regularization played a crucial role in stabilizing the model and ensuring meaningful coefficient values. Future work could explore other regularization techniques like Lasso Regression or more advanced models to further improve prediction accuracy.