Exploring and Modeling the Boston Housing Dataset

Donal Moloney

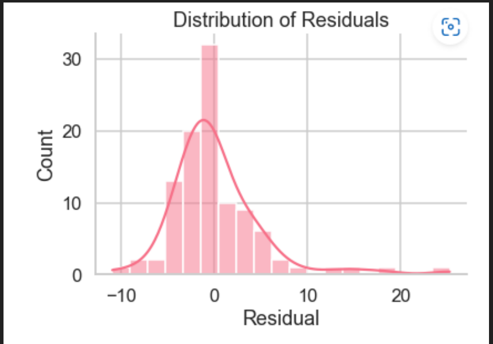
Introduction

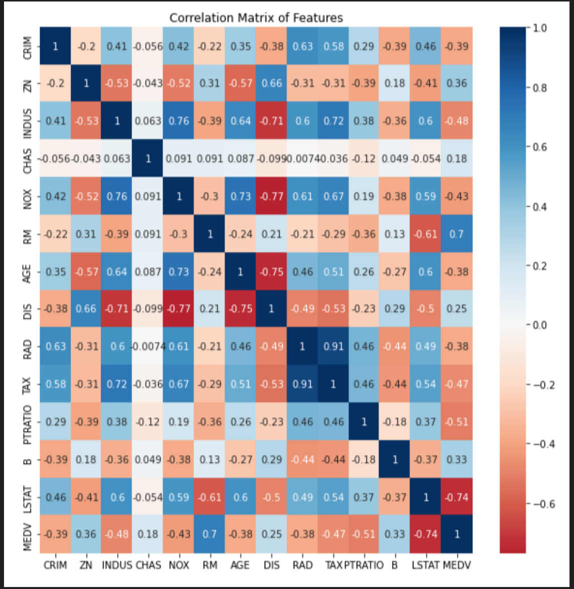
- A detailed overview of the Boston Housing dataset. This compilation encompasses housing price data from Boston's suburban areas and integrates various socioeconomic factors that may influence these costs
- Project Objectives: Explore and visualize the Boston Housing dataset, create a linear regression model for housing value prediction, and evaluate its performance.

Exploring the DataSet

- Through a histogram of the target variable, the distribution of housing prices (MEDV) appears largely normal with some outliers at the higher end.

- By employing a heatmap of the correlation matrix and scatterplots, we observe distinct positive and negative correlations among features, highlighting key determinants for housing value predictions.





Exploring the DataSet

Data Visualization Insights:
The average number of rooms per dwelling (RM) shows a robust positive correlation with housing prices, while the percentage of lower status population (LSTAT) inversely correlates. These two features likely serve as key predictors of housing prices.

SICUT PATRIBUS SIT DEUS NOBIS.
BOSTONIA
CONDITA AD
·1630·
CIVITATIS REGIMINE DONATA AD 1822

Data Partitioning for Model Development & Assessment:

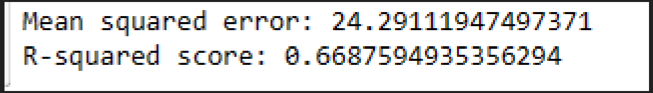
- Divided the data into training and testing subsets to foster a robust foundation for our linear regression model.

- Leveraging the Sklearn library, the training set was harnessed to construct a model that predicts housing values derived from the dataset's various features.

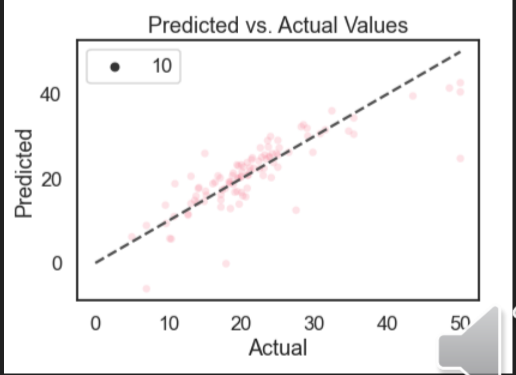
Data Partitioning for Model Development & Assessment pt 2:

- Utilizing the testing set, we gauged our model's efficacy through the Mean Squared Error (MSE) and R-squared score. The MSE reflects the precision of our housing price predictions, while the R-squared score demonstrates the model's fit to the data.

- Model performance was visualized using scatterplots of predicted versus actual values and the distribution of residuals. Both the scatterplot and the residual distribution confirm the model's commendable accuracy in predicting housing prices.



```
Mean squared error: 24.291119474973711
R-squared score: 0.6687594935356294
```



Techniques to Improve

- Incorporating cross-validation enhances the model's predictive performance. As a technique to measure model efficacy and deter overfitting, cross-validation ensures a more accurate forecast on unseen data.

- Incorporating regularization strategies helps mitigate overfitting. Techniques such as Lasso and Ridge regression, which penalize substantial coefficients, contribute to model robustness, bolstering its generalization capability on unseen data.

Techniques to Improve Part 2

- Exploring Non-linear Relationships: While the current linear regression model presumes a linear relation between the target variable and features, potential non-linear associations may exist. Delving into these non-linear connections could enhance the model's accuracy.

- Considering Alternative Modeling Techniques: While linear regression serves as one predictive method for housing values, exploring other approaches like neural networks, decision trees, and random forests may yield even more precise results.

Boston Housing Dataset Project Summary:

- In this project, we investigated and modeled the Boston Housing dataset by analyzing the performance of a linear regression model, visualizing the distribution of the objective variable, and relationships between features. We discovered that significant predictors of housing prices include the average number of rooms per dwelling and the percentage of the people with lower socioeconomic standing. Our model's MSE was 24.29 and its R-squared score was 0.67, which shows that it can forecast housing prices reasonably accurately but could still use some work.

- In order to enhance the performance of our model, we proposed a number of methods, including the addition of cross-validation, regularization strategies, the investigation of non-linear relationships, and the use of alternative modeling methods.

- Overall, this project offered insights into the variables that affect housing prices and how we can forecast them, and it established a solid framework for further research in this field.