Reflective Report on Portfolio Part4

Sayedali Mohseni - Student ID: 45765778

2023-10-22

The Reflective Report of Portfolo Part4

The Choice of the Dataset

The choice of the dataset in my portfolio, the Boston Housing dataset, was influenced by several factors. Firstly, it's a classic dataset widely employed for educational purposes in machine learning. Its combination of numerical and categorical features, along with its well-defined target variable (median home value), makes it suitable for teaching various machine learning techniques. Secondly, it reflects a real-world problem: predicting housing prices, which is of practical interest to homeowners, real estate agents, and potential buyers. Lastly, the dataset is readily accessible and well-documented, making it suitable for educational purposes.

The problem Identification

The problem of predicting the median value of owner-occupied homes in Boston was identified based on the dataset's nature and objectives. The dataset's target variable, 'MEDV,' naturally led to the problem statement. Additionally, the goal of predicting housing prices is a common use case for regression analysis and machine learning, making it relevant and instructional. This problem provides a valuable opportunity to apply regression modeling techniques and gain insights into the factors influencing housing prices.

Selection of the Machine Learning Models Applied in the Portfolio

In my portfolio, I used two primary machine learning models: Simple Linear Regression, Polynomial Regression and K-means Clustering. These choices were made for the following reasons:

- **Simple Linear Regression**: It serves as a foundational model for regression problems and helps establish a baseline. Its simplicity makes it easy to understand and interpret. It's suitable for cases where there's a linear relationship between the features and the target variable.
- **Polynomial Regression**: I introduced Polynomial Regression to capture more complex relationships that might exist between the features and the target variable. It's a suitable choice when linear relationships are not sufficient. However, it's essential to be cautious about overfitting when using polynomial regression.

Both models are appropriate for solving the problem of predicting housing prices because they can capture the potential linear and non-linear dependencies between the input features (e.g., crime rate, number of rooms) and the target variable (median home value). By using these models, I aimed to explore and compare their predictive performance and understand how well they fit the data.

- **K-means Clustering**: I ignored the labels in the data and tried to apply K-means Clustering which is considered as classification model and a unsupervised machine learning method to explore if there is any instinct clusters in the data, and any other relationship other than finding from the regression methods.

Conclusions

The insights and conclusions drawn from the study are as follows:

- **Simple Linear Regression**: This model provided a baseline understanding of the relationship between individual features and the median home value. The coefficients of the features indicated their influence. For example, 'RM' (average number of rooms) had the highest positive coefficient, suggesting that more rooms lead to higher home values. This aligns with the intuitive expectation that larger homes are generally more expensive.
- **Polynomial Regression**: The polynomial regression model improved predictive performance compared to simple linear regression. It captured more complex relationships between features and the target variable. However, the unusually high Adjusted R^2 value raised concerns about potential overfitting. This result suggests that while polynomial regression can enhance predictions, it's essential to address overfitting to ensure the model's generalizability.

The results were consistent with intuitive expectations, such as the positive correlation between the number of rooms and home values. However, the overfitting issue in polynomial regression highlights the importance of model evaluation and potential regularization techniques to enhance model robustness. - **K-means Clustering**: The model has been applied focusing on two features, "Average Number of Rooms per Dwelling (RM)" and "Percentage of Lower Status of the Population (LSTAT). Also, it found three different custers.

This clustering suggests that areas with more rooms per dwelling tend to have a lower percentage of the lower-status population, and vice versa. The centroids provide a good representation of the average characteristics of each cluster.

These insights and conclusions help in understanding the factors influencing housing prices in Boston and the trade-offs between model complexity and performance.