BOSTON HOUSING DATASET ANALYSIS REPORT

4-29-2024

MATTHEW HENAO

Z23685608
Professor Juan Yepes

Overview

The Boston Housing Dataset is a well-known dataset used primarily in machine learning and statistics for predicting housing prices through regression analysis. It was originally compiled by the U.S. Census Service in the 1970s and has been extensively used to analyze and predict housing market behaviors.

Data Description

These features are both directly and indirectly related to the housing prices in the area. Below is a description of each feature included in the dataset:

- 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
- NOX: Nitric oxides concentration (parts per 10 million).
- RM: Average number of rooms per dwelling.
- AGE: Proportion of owner-occupied units built prior to 1940. (target variable)
- 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: where B is the proportion of black residents by town.
- LSTAT: Percentage of lower status of the population. (target variable)
- MEDV: Median value of owner-occupied homes in $1000s

VIF Analysis

Constant (VIF = 641.74): The high VIF for the constant term typically reflects the inclusion of other constant terms in the model or a high degree of multicollinearity among the variables.

CRIM (VIF = 1.83) and CHAS (VIF = 1.09): These variables exhibit low VIFs, indicating that they do not have strong linear relationships with other predictors in the model. This suggests that these variables provide unique information not captured by other variables.

ZN (VIF = 2.32), INDUS (VIF = 3.99), RM (VIF = 2.18), PTRATIO (VIF = 1.98), B (VIF = 1.38), and LSTAT (VIF = 3.24): These variables have moderate VIF values, suggesting moderate correlation with other variables but not enough to be overly concerning according to the VIF threshold of 10.

NOX (VIF = 4.28), DIS (VIF = 4.06), and MEDV (VIF = 3.86): These values are approaching the higher end of the moderate range. They indicate some level of multicollinearity, but still below the threshold that would typically cause alarm.

RAD (VIF = 7.76) and TAX (VIF = 9.20): These variables are approaching the threshold of 10, suggesting that they are highly correlated with other variables in the model.

Analysis of Linear Regression Model Performance

Coefficients & Intercept:

The coefficients derived from the model for most variables are extremely close to zero, with the exception of a significant coefficient for one of the variables (approximately 7.10). This suggests that most of the predictors have minimal influence on the model, except for this particular variable which appears to be a primary driver in predicting the target variable.

The model's intercept is approximately 12.46, indicating the average expected value of the dependent variable when all predictors are at their mean value.

Mean Squared Error (MSE):

The MSE is extremely low (approximately 2.15×10−28 2.15×10 −28), indicating that the model predictions are almost exactly the same as the actual values. In practical scenarios, such a low MSE is unusual and might indicate overfitting or issues with the test dataset or model setup.

R-squared (R^2):

The R^2 value is 1.0, which means the model explains 100% of the variance in the dependent variable from the predictors. This is an exceptionally perfect score and is typically suspect in real-world data analyses as it suggests a potentially overfitted model.

Analysis of Classification Model Performance Using LDA

The model achieved an overall accuracy of 86.27%. This indicates that, on average, the model correctly predicts the class of an observation 86.27% of the time across all predictions made.

Detailed Classification Metrics

Class 0 (Precision: 1.00, Recall: 0.67, F1-Score: 0.80): The model perfectly identifies all relevant instances of Class 0 (precision = 1.00), but it only correctly identifies about 67% of actual instances of this class (recall = 0.67). This suggests some misclassification of Class 0 instances as belonging to other classes.

Class 1 (Precision: 0.80, Recall: 0.70, F1-Score: 0.74): For Class 1, the model is quite precise but not as robust in recall, indicating that while the predictions made are often correct, it misses about 30% of actual cases.

Class 2 (Precision: 0.68, Recall: 0.85, F1-Score: 0.76): Class 2 shows a lower precision but higher recall, meaning the model captures a good portion of Class 2 instances but also mislabels other classes as Class 2.

Class 3 (Precision: 0.96, Recall: 0.98, F1-Score: 0.97): The model performs excellently with Class 3, with both high precision and recall, indicating both accurate and comprehensive capture of Class 3 instances.

Analysis of Polynomial Regression Model Performance

Model Descriptions and Performance

3rd-degree Polynomial Regression (MSE: 36.52):

The model with polynomial features of degree 3 resulted in an MSE of 36.52. This indicates a moderate level of prediction error, suggesting that while the model can capture some non-linearity in the data, there might be room for improvement or a need for further tuning.

4th-degree Polynomial Regression (MSE: 19.68):

Increasing the degree to 4 significantly improves the model's performance, with the MSE reducing to 19.68. This improvement suggests that the additional complexity provided by the 4th-degree terms helps the model to better capture the underlying patterns in the data.

5th-degree Polynomial Regression (MSE: 35.87):

Surprisingly, further increasing the polynomial degree to 5 results in an MSE that is nearly as high as the 3rd-degree model, at 35.87. This could indicate overfitting, where the model becomes too tailored to the training data, losing its generalizability and thus performing poorly on new, unseen data.

Analysis of Decision Tree Regressor Performance on the MEDV Variable

Overall Accuracy: The classifier achieves an accuracy of 79.41%. This metric indicates a good general performance but hides class-specific details.

Class-specific Performance:

High (Precision: 1.00, Recall: 0.25, F1-Score: 0.40): Perfect precision indicates that all predictions of the 'High' category were correct, but the low recall shows that the model failed to identify most of the actual 'High' cases.

Low (Precision: 0.81, Recall: 0.89, F1-Score: 0.85): The model performs well for the 'Low' category, correctly identifying a high percentage of cases, though there are some false positives.

Medium (Precision: 0.74, Recall: 0.68, F1-Score: 0.71): This category sees balanced but moderate scores in both precision and recall, indicating a fair performance.

K-Means Clustering and Anomaly Detection Analysis

Cluster Profiling

Cluster 0: Characterized by relatively high values in features like 'nox', 'age', and 'tax', suggesting this cluster might consist of older areas with higher pollution and tax rates.

Cluster 1: Shows moderate negative values in 'indus' and 'nox' and positive values in 'zn' and 'rm', indicating this cluster likely represents residential areas with larger homes and less industrial activity.

Cluster 2: Marked by extreme negative values in several features and high values in 'MEDV_category_High', suggesting these areas are significantly different, potentially more affluent or economically distinct from others.

Feature Importance

The absolute values of the centroids indicate the relative importance of each feature in defining the cluster. Higher values indicate a stronger role in the cluster's profile. For instance, 'MEDV_category_High' has a dominant presence in Cluster 2, reflecting its unique demographic or economic status.

Cluster Validation Metrics

Silhouette Score (0.328): Indicates a fair separation between clusters, but there's room for improvement as values closer to 1 represent better-defined clusters.

Davies-Bouldin Score (1.025): A lower score (closer to 0) is better, suggesting that the clusters are reasonably compact and well-separated, although there could be some overlap.

Anomaly Detection

Detected Anomalies: A total of 21 data points were identified as anomalies based on their distance from the nearest cluster centroid. These are likely extreme or unusual cases within the dataset.

Anomalies Profile: These anomalies include properties with unusually high or low values in certain features such as 'crim', 'zn', 'indus', 'nox', and 'rm'. For example, index 388 shows extremely high crime rates and pollution levels, which significantly differ from most other data points.

Analysis of Linear Regression Performance with Cross-Validation

Cross-Validation Results

Scores for Each Fold: The Mean Squared Error (MSE) for each fold are as follows:

Fold 1: 4.580

Fold 2: 4.568

Fold 3: 5.107

Fold 4: 4.407

Fold 5: 4.995

These values represent the model's error metric in each cross-validation fold, indicating the average squared difference between the predicted and actual values.

Mean Cross-Validation Score: The mean MSE across all folds is approximately 4.731. This average indicates the model's typical performance in predicting 'LSTAT' across different subsets of the data, providing a robust measure of its predictive accuracy.

Standard Deviation of Cross-Validation Scores: The standard deviation of the MSE scores is about 0.270.