COE 379L Lab Report

Parth Patki

3/11/2025

pap2389

Part 1 Comments

The dataset contains 20,634 entries and 9 columns, including Median Income (MedInc), House Age (HouseAge), Average Rooms (AveRooms), Average Bedrooms (AveBedrms), Population, Average Occupants per Household (AveOccup), Latitude, Longitude, and the target variable price_above_median, which shows whether a house's price is above the median.

The statistical summary reveals several insights. Median Income ranges from 0.5 to 15.0, averaging 3.87, indicating a slight skew towards lower-income levels. House Age spans from 1 to 52 years, with an average of 28.64 years. The maximum values for Average Rooms (141.91) and Average Bedrooms (34.06) are notably higher than their means, suggesting the presence of outliers. Population varies significantly, with a maximum of 35,682, hinting at some densely populated areas. Average Occupants per Household peaks at 1,243, another clear outlier to consider. Latitude and Longitude values are consistent with California's geographic range. The target variable, price_above_median, appears balanced, with a mean of 0.5 and a standard deviation close to 0.5, indicating an even distribution of homes above and below the median price.

Histograms from the univariate analysis highlighted several key trends. Median Income shows a skew towards lower values, with a few outliers in the higher-income range. House Age is primarily concentrated between 20-40 years. Both Average Rooms and Average Bedrooms have long-tailed distributions, confirming the presence of extreme outliers. Population and Average Occupants per Household are also highly skewed, indicating that some blocks have a very high density. Latitude and Longitude distributions align with California's geographic boundaries.

Part 3

Model Training Techniques

To predict whether a house's price is above the median, a variety of classification models were utilized, including K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and AdaBoost. Each model was trained using the dataset with price_above_median as the target variable. To ensure the models could generalize effectively, the dataset was split into separate training and testing sets. This approach helps evaluate how well the models perform on unseen data. Additionally, standardization was applied to features where necessary—particularly for distance-based models like KNN—to ensure that all features were on a similar scale, thereby improving model accuracy and performance.

Model Optimization Techniques

To enhance model performance, hyperparameter tuning was conducted for each classification model. For K-Nearest Neighbors (KNN), various values of k were tested to find an optimal balance between bias and variance. The Decision Tree model was fine-tuned by adjusting the maximum depth, helping to prevent overfitting while maintaining predictive accuracy. Random Forest, being an ensemble method, was optimized by experimenting with the number of trees and the maximum number of features considered at each split to identify the best-performing combination. For AdaBoost, a base Decision Tree classifier was used, and the number of boosting rounds was carefully adjusted to maximize performance while ensuring the model maintained good generalization on unseen data.

Model Performance Comparison

Among the evaluated models, Random Forest achieved the highest accuracy at 88.66%, followed by AdaBoost at 84.78%, Decision Tree at 84.01%, and KNN at 82.07%. Random Forest not only led in overall accuracy but also demonstrated the highest precision and recall across both classes, highlighting its strength in handling complex decision boundaries effectively. The confusion matrices further supported these findings, showing that Random Forest and AdaBoost resulted in fewer misclassifications compared to the other models. In contrast, KNN faced slight challenges, likely due to its dependence on distance metrics, which can be sensitive to outliers and variations in feature scaling.

Recommended Model

Based on the performance metrics, Random Forest emerges as the most suitable model for this dataset. It provides a strong balance between precision and recall, effectively managing outliers and capturing complex feature interactions. Its ensemble approach, which combines multiple decision trees, helps reduce the risk of overfitting compared to using a single Decision Tree. This makes Random Forest not only the most accurate but also the most reliable option for predicting whether a house's price is above the median.

Most Important Metric

For this dataset, both accuracy and recall are key performance metrics. Given that the dataset is balanced, with roughly 50% of houses priced above the median, overall accuracy serves as a reliable measure of model performance. However, recall holds particular importance, as misclassifying high-value properties as low-value (false negatives) could lead to poor decision-making in real estate pricing. Accurately identifying high-priced houses is essential to avoid undervaluation. While precision is also valuable, prioritizing recall ensures that a greater number of high-priced properties are correctly classified, reducing the risk of overlooking valuable assets.