Alireza mohammadshafie group project.

1. Introduction

The goal of this project is to accurately predict housing prices in California using machine learning techniques. By leveraging a variety of features that describe each property and its surroundings, we aim to develop a model that generalizes well and can provide reliable price estimations. Among several regression methods tested, Random Forest Regression demonstrated superior performance and was chosen as the final model.

2. Data Description and Preprocessing

We used the California Housing Dataset, which includes various features such as median income, housing median age, average rooms, and more. The dataset was carefully explored for missing values and outliers; none significant were found that required removal. Data visualization provided insights into feature correlations and distribution patterns, which helped guide feature selection.

3. Feature Engineering

To enhance the model's predictive power, we created a new feature called RoomDensity, calculated as the ratio of average rooms to average occupants per household.This feature captures the spatial density of housing units, providing insight into how roomy or crowded a household is. Intuitively, a higher room density may correlate with higher housing prices due to increased comfort and desirability. Incorporating this engineered feature allowed the model to better understand nuances in the housing data beyond the original variables.

All features—including the new RoomDensity—were scaled before model training.

4. Model Building and Evaluation

The dataset was split into training and test sets to evaluate model generalization. Several regression algorithms were tested, including Linear Regression, Support Vector Regression, K-Nearest Neighbors, and ensemble methods like Random Forest and Gradient Boosting. Evaluation metrics included Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared.

Among tested approaches, Random Forest Regression demonstrated the best balance of accuracy and interpretability, with an R-squared of approximately 0.81 on the test set, indicating that it explains about 81% of the variance in housing prices.

5. Model Tuning and Selection

To optimize our Random Forest model, we performed hyperparameter tuning using GridSearchCV. The grid explored the following parameters:

- Number of trees (n_estimators): 50, 100, 200
- Maximum tree depth (max_depth): 10, 25, 30, None (no limit)
- Maximum number of features considered at each split (max_features): 'auto', 'sqrt'

Using 5-fold cross-validation, the grid search identified the best combination of hyperparameters that minimized the Mean Squared Error (MSE) on the validation sets. The tuned model slightly improved performance compared to default settings, This disciplined search contributed to a better trade-off between model complexity and generalization, leading to more accurate predictions.

6. Results and Analysis

After tuning, the Random Forest Regression model achieved the following results on the test data:

- Test MSE: 0.245
- Test RMSE: 0.495
- Test MAE: 0.326
- Test R-squared: 0.813

These improvements reflect a more precise fit to the housing price data, with the model explaining approximately 81% of the variance. Our custom feature, RoomDensity, played a role in capturing household spatial characteristics, enhancing the model's understanding of the data.

Visualizations confirm the predicted median house values closely follow the actual values, reinforcing the model’s effectiveness.

7. Conclusion

This project successfully applied Random Forest Regression, combined with custom feature engineering and hyperparameter tuning, to achieve a reliable housing price predictor.

References

1-https://scikit-learn.org/stable/supervised_learning.html

2-https://www.geeksforgeeks.org/regression-in-machine-learning/

3-https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/