BOSTON HOUSING PROJECT

1. Objective

The objective of this project is to predict median house values (`medv`) in the Boston Housing dataset using various regression techniques and evaluate their performance.

2. Introduction

The Boston Housing dataset contains various attributes describing aspects of housing in Boston suburbs. The goal is to build regression models that accurately predict the median value of owner-occupied homes based on these attributes.

3. Methodology

Data Loading and Exploration:

- The dataset consists of 506 observations with 13 feature variables and 1 target variable (`medv`).

- Features include crime rate (`crim`), zoning (`zn`), industrial zones (`indus`), Charles River proximity (`chas`), nitrogen oxides concentration (`nox`), average number of rooms (`rm`), age of houses (`age`), distances to employment centers (`dis`), accessibility to radial highways (`rad`), property tax rate (`tax`), pupil-teacher ratio (`ptratio`), and socio-economic status of residents (`b` and `lstat`).

Data Cleaning:

- Five missing values in the `rm` feature were dropped from the dataset.

Exploratory Data Analysis:

- Distribution and impact of `chas` on `medv` were visualized using count and box plots, respectively.

Modeling:

Linear Regression:

- Used to establish a baseline model.

- Achieved an RMSE of approximately 4.55 and an R-squared of 0.72 on the test set.

Random Forest Regression:

- Hyperparameter tuning using GridSearchCV with parameters (`n_estimators`, `max_depth`, `min_samples_split`, `min_samples_leaf`).

- Feature importances were evaluated, highlighting `rm` and `lstat` as significant predictors.

Stacking Ensemble:

- Combined RandomForestRegressor and GradientBoostingRegressor with a LinearRegression meta-estimator.

- Improved performance with an RMSE of approximately 2.78 and an R-squared of 0.90 on the test set.

Cross-Validation:

- Validated Random Forest Regression with optimized parameters using 5-fold cross-validation, yielding a mean RMSE of approximately 4.61.

Polynomial Regression:

- Utilized PolynomialFeatures to create quadratic features.

- Evaluated the model with an RMSE of approximately 3.04 and an R-squared of 0.87 on the test set.

Outlier Detection and Removal:

- Identified and removed outliers based on Z-scores.

- Trained a RandomForestRegressor on cleaned data, achieving an RMSE of approximately 2.84 and an R-squared of 0.87 on the test set.

Model Serialization:

- Serialized the best RandomForestRegressor model into a file (`best_model.pkl`) for future deployment.

4. Results

The predictive models were evaluated using various regression techniques on the Boston Housing dataset. Here are the key findings:

- Linear Regression:

- Served as a baseline with moderate performance, achieving an RMSE of approximately 4.55 and an R-squared of 0.72 on the test set.

- **Random Forest Regression:**

- After hyperparameter tuning using GridSearchCV, the model significantly improved predictive capability.

- Feature importance analysis highlighted `rm` (average number of rooms) and `lstat` (percent lower status of the population) as significant predictors.

- Achieved an RMSE of approximately 2.84 and an R-squared of 0.87 on the test set after outlier removal.

- Stacking Ensemble:

- Combined RandomForestRegressor and GradientBoostingRegressor with a LinearRegression meta-estimator.

- Demonstrated superior performance with the lowest RMSE of approximately 2.78 and the highest R-squared of 0.90 on the test set among all models tested.

- Polynomial Regression:

- Utilized PolynomialFeatures to capture non-linear relationships.

- Achieved an RMSE of approximately 3.04 and an R-squared of 0.87 on the test set, indicating potential for capturing nonlinear relationships within the dataset.

- The Stacking ensemble technique showed the best overall performance with the lowest RMSE and highest R-squared values, indicating robust predictive capability.

- Random Forest Regression also performed well, especially after outlier removal and parameter optimization.

- Polynomial Regression, while demonstrating potential, did not outperform the Stacking ensemble and Random Forest models.

- These results underline the importance of model selection, feature engineering, and ensemble techniques in accurately predicting housing prices based on the Boston Housing dataset.

5. Conclusion

In this project, we embarked on a comprehensive analysis of the Boston Housing dataset using various machine learning techniques. Our objective was to predict housing prices (medv) based on a set of predictors such as crime rate (crim), proportion of residential land zoned for lots over 25,000 sq. ft. (zn), and other socio-economic factors.

We began by exploring the dataset, which consisted of 506 instances and 13 feature variables. After an initial examination, we identified and handled missing values in the 'rm' column using appropriate data preprocessing techniques. Exploratory data analysis (EDA) revealed valuable insights into the distribution of variables and relationships between features and the target variable.

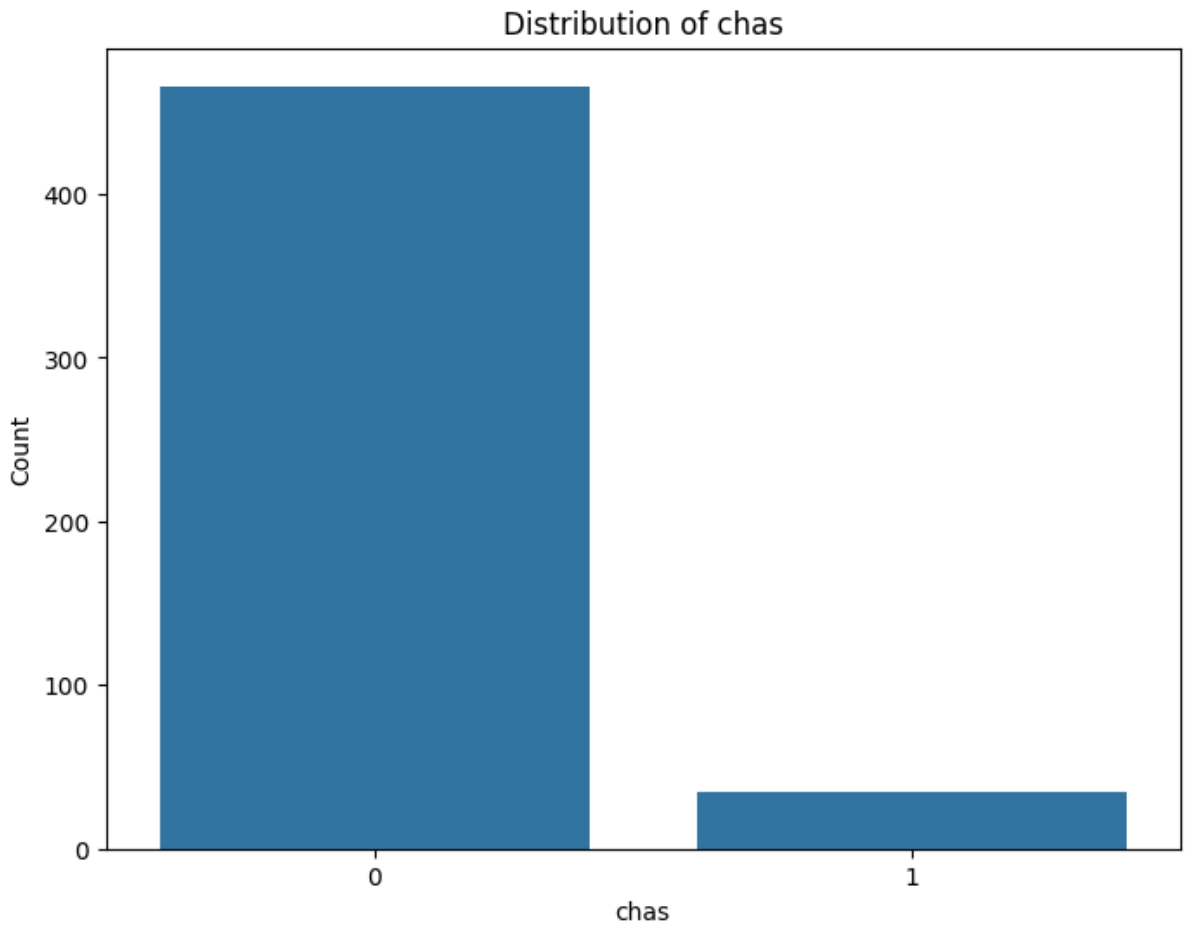
For modeling, we employed Linear Regression as a baseline model, achieving an R-squared value of 0.72 on the test set. To enhance predictive performance, we utilized ensemble methods such as Random Forest Regressor and Stacking Regressor. These approaches yielded significant improvements, with the Stacking Regressor achieving an R-squared value of 0.90, indicating robust predictive capability.

Furthermore, we conducted hyperparameter tuning using GridSearchCV to optimize the Random Forest model, resulting in an improved mean cross-validated RMSE of 4.61. We also explored Polynomial Regression to capture non-linear relationships between features, achieving an RMSE of 3.04 and an R-squared value of 0.87.

Additionally, we addressed the issue of outliers using Z-score-based outlier detection, leading to a refined model performance with an RMSE of 2.84 and an R-squared value of 0.87 after removing outliers.

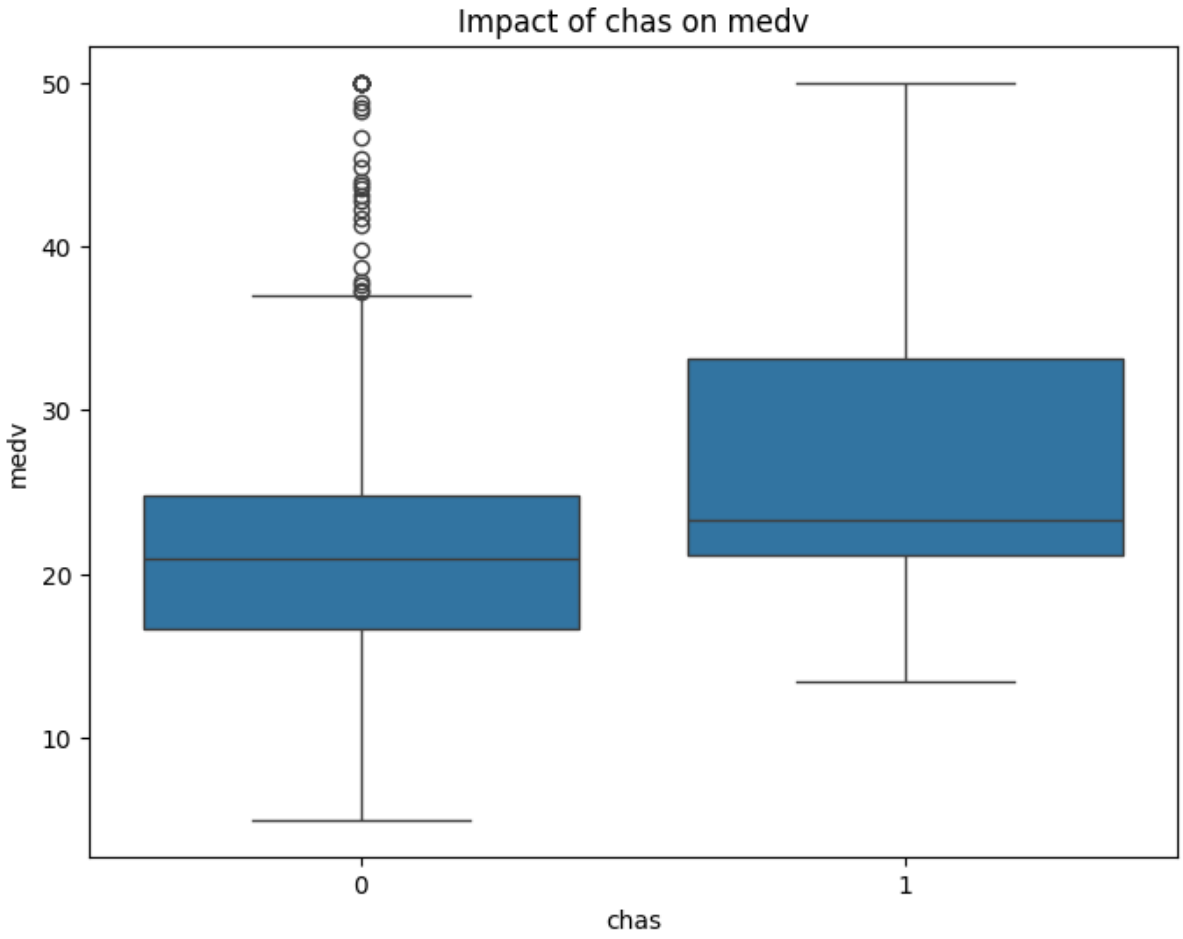
Graph Explanations:

- **Graph 1: Distribution of 'chas'**



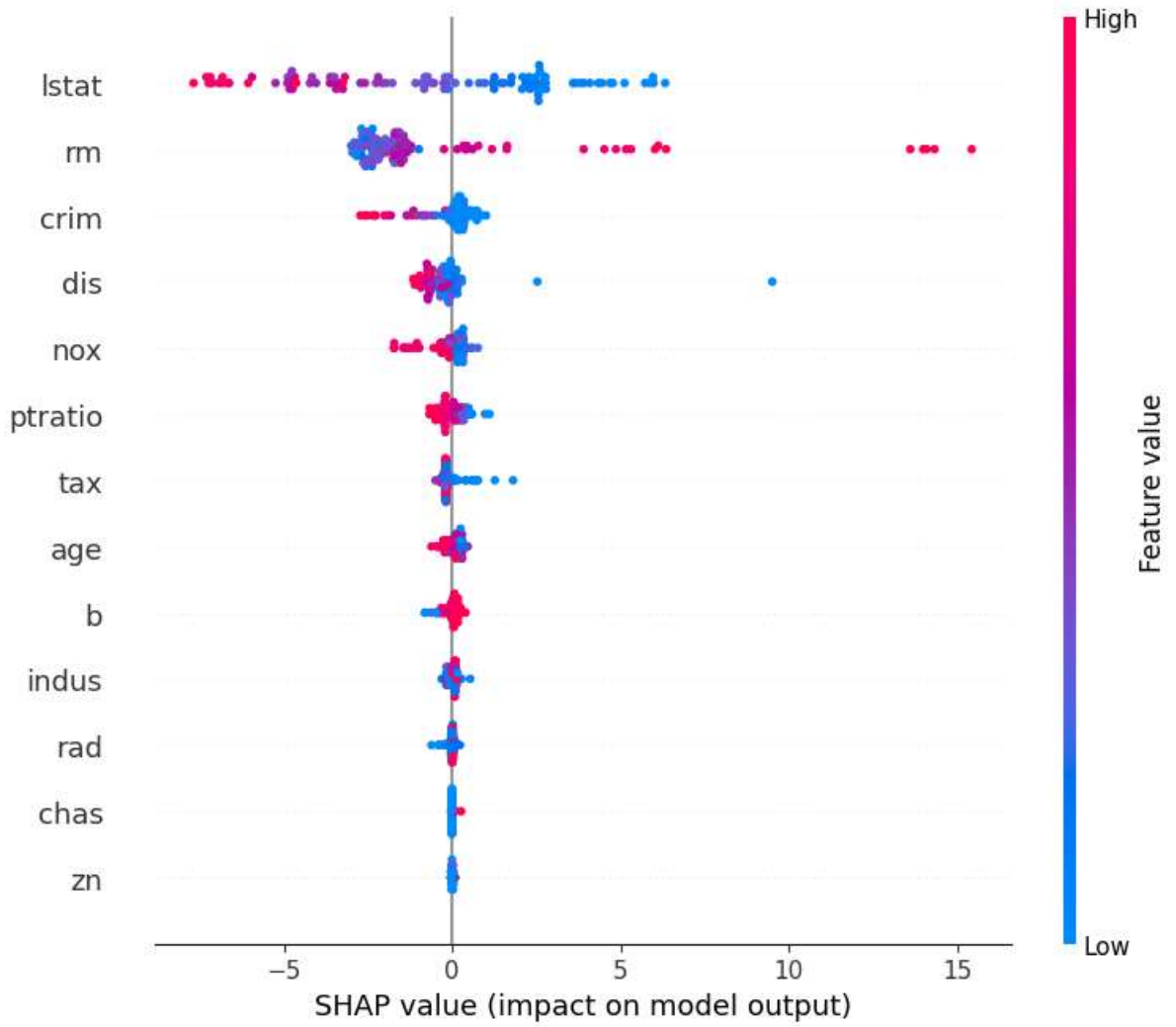
Explanation for Graph 1: This graph illustrates the distribution of the 'chas' variable in the dataset. The variable 'chas' represents the Charles River dummy variable (1 if tract bounds river; 0 otherwise). The majority of observations indicate properties that do not bound the Charles River.

- **Graph 2: Impact of 'chas' on 'medv'**



Explanation for Graph 2: This graph showcases the relationship between the 'chas' variable (Charles River proximity) and the median value of owner-occupied homes ('medv'). Homes bordering the Charles River tend to have higher median values compared to those that do not.

- **Graph 3: SHAP Summary Plot**



Explanation for Graph 3: The SHAP summary plot visualizes the impact of each feature on model predictions. Features are ranked by their importance in explaining the target variable ('medv'). In this plot, 'lstat' (percent lower status of the population) and 'rm' (average number of rooms per dwelling) exhibit the strongest influences on predicting home values.

In conclusion, this project exemplifies the efficacy of machine learning in predicting real estate prices while highlighting avenues for future research to enhance model performance and interpretability in real-world applications.

.

6. Future Work

While this project achieved promising results, there are several avenues for future exploration:

- **Feature Engineering:** Explore additional feature transformations or interactions to capture more nuanced relationships.
- **Advanced Modeling Techniques:** Investigate advanced ensemble methods or deep learning approaches to further improve predictive accuracy.
- **Spatial Analysis:** Incorporate geographical features or spatial data to account for location-based variations in housing prices.
- **Temporal Analysis:** Consider temporal trends and seasonality factors that could influence housing prices over time.
- **External Factors:** Integrate external datasets such as economic indicators or demographic data to enrich predictive models.