Boston Housing Price Prediction - Final Report

Name: Samad Mehboob

Email: samadmehboob940@gmail.com

1. Project Objective

The goal of this project is to train and evaluate multiple regression models on the Boston Housing dataset to predict house prices. The process involves exploratory data analysis, feature engineering, model evaluation, hyperparameter tuning, and model comparison to identify the most effective predictive model.

2. Data Exploration

The dataset contains 506 samples and 14 features. The target variable is 'PRICE' (in $1000s). Initial exploration reveals a right-skewed distribution for price and strong correlations with the following features:

- RM (average rooms per dwelling): +0.70
- LSTAT (lower status of population): -0.74
- PTRATIO (pupil-teacher ratio): -0.51

Missing values are not present, and distributions and relationships were visualized using histograms and scatterplots.

3. Feature Engineering

New features such as ROOMSPERHOUSE, TAXPERROOM, and NOXRM were created. Skewed features (e.g., CRIM, LSTAT, NOX, DIS) were log-transformed to reduce skewness. Selected features include RM, LSTAT, PTRATIO, INDUS, TAX, CRIM, and engineered/log-transformed features.

4. Model Evaluation & Performance

The following models were trained and evaluated using Mean Squared Error (MSE) and R-squared metrics:

- Linear Regression: $R^2 = 0.73$
- Lasso Regression: $R^2 = 0.72$
- Ridge Regression: $R^2 = 0.74$
- Random Forest: $R^2 = 0.86$
- Tuned Random Forest: $R^2 = 0.88$ (Best)

The tuned Random Forest Regressor was the top-performing model with optimized hyperparameters:

* n_estimators = 200
* max_depth = 10
* min_samples_split = 5

5. Key Feature Importance

Based on the Random Forest model, the most important features contributing to house price prediction are:

1. RM (average rooms per dwelling)
2. LSTAT (lower status population)
3. PTRATIO (pupil-teacher ratio)
4. LOG_LSTAT (log-transformed LSTAT)
5. LOG_CRIM (log-transformed crime rate)

6. Recommendations

1. Use the saved Random Forest model (`boston_housing_rf_model.pkl`) for accurate predictions.
2. Focus on improving or collecting more data on key features like RM and LSTAT.
3. Consider using ensemble models and tuning them further with more granular parameter grids.
4. For real deployment, integrate the model into an interactive dashboard or web app.

7. Conclusion

This project successfully built a predictive model for Boston housing prices using the Random Forest algorithm. After comparing various models and tuning, the best model was saved for future use. The approach demonstrates the effectiveness of ensemble methods in regression problems.