Predicting Boston House Prices Using Machine Learning

By: Menna Mahmoud EL-Bagoury

Problem Statement

The Boston housing market is affected by numerous factors, including crime rates, residential area proportions, property age, and room numbers, making price prediction a complex challenge. Buyers, sellers, and real estate agents require accurate pricing predictions to make informed decisions. This project aims to address this challenge by developing a machine learning model that can predict housing prices in Boston with high accuracy. This solution would improve market efficiency, ensure fair valuation, and support better decision-making in the real estate market.

Objectives

1. **Develop a machine learning model** to predict housing prices based on features like crime rate, number of rooms, and age of the property.

2. **Identify key factors** that most influence housing prices in the Boston area.

3. **Achieve high predictive accuracy**, improving on traditional methods through data-driven insights.

Data Description

Source: The dataset used is the Boston Housing dataset, widely known for its application in regression tasks.

Dataset Details:

Training dataset: 506 records with 14 columns, including the target variable `MEDV` (median house price).

Features:

- **CRIM:** Crime rate by town
- **ZN:** Proportion of residential land zoned for large lots
- **INDUS:** Proportion of non-retail business acres per town
- **CHAS:** Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- **NOX:** Nitric oxide concentration (parts per 10 million)
- **RM:** Average number of rooms per dwelling
- **AGE:** Proportion of owner-occupied units built before 1940
- **DIS:** Weighted distances to five Boston employment centers
- **RAD:** Index of accessibility to radial highways
- **TAX:** Property tax rate per $10,000
- **PTRATIO:** Pupil-teacher ratio by town
- **B:** Proportion of Black population by town
- **LSTAT**: Percentage of lower status of the population
- **MEDV**: Median value of owner-occupied homes (target variable)

Missing Values: This dataset is relatively clean but still required preprocessing for optimal model performance.

Abstract

This project employs various machine learning techniques to predict Boston housing prices, addressing the need for accurate and fair price estimation. Using a structured machine learning pipeline, data preprocessing techniques were applied, and models like Linear Regression, Decision Trees, and Random Forests were trained and tuned. The Random Forest model, after optimization, achieved the highest accuracy, demonstrating its effectiveness in predicting

housing prices. The results showcase the importance of feature engineering and hyper parameter tuning in achieving reliable predictions.

Methodology

1. Data Collection:

- The dataset was sourced from the Boston Housing data repository, containing 506 records with 13 features and one target variable (house price).

2. Data Cleaning:

- Outliers and influential data points were detected and addressed.

- Missing values were minimal but handled where necessary to avoid biased predictions.

3. Exploratory Data Analysis (EDA):

- Conducted a detailed EDA to understand the distribution and relationships among variables.

- Visualized feature distributions and correlations to understand the influence of variables like *RM, LSTAT*, and *PTRATIO* on *MEDV*.

- High correlations were observed between *LSTAT (*percentage of lower status of the population) and *MEDV*, as well as *RM* (average number of rooms per dwelling) and *MEDV*.

4. Feature Engineering:

- Created interaction features and polynomial features to capture non-linear relationships.

- Applied feature scaling using StandardScaler for continuous features to normalize their distribution.

- Categorical variables (such as *CHAS*) were one-hot encoded to make them suitable for machine learning models.

5. Model Selection:

- Tried various models, including:

- Linear Regression: Baseline model to observe linear relationships.

- Decision Tree Regressor: Allowed capturing non-linearities in the data.

- **Random Forest Regressor:** Combined multiple trees for better generalization and accuracy.

6. Hyperparameter Tuning:

- Tuned models using GridSearchCV and RandomizedSearchCV to optimize parameters such as *max_depth, min_samples_split*, and *n_estimators* for Random Forest.

- Achieved the best performance with Random Forest, which showed a well-balanced trade-off between bias and variance.

7. Evaluation Metrics:

- RMSE (Root Mean Squared Error): Measured average error between predicted and actual prices.

- R-squared: Provided insight into the proportion of variance explained by the model.

- The Random Forest model achieved the lowest RMSE and highest R-squared, indicating strong predictive performance.

Findings

1. Feature Importance:

- Variables like *RM, LSTAT*, and *PTRATIO* were identified as the most significant predictors of housing prices.

- *RM* (average number of rooms) positively correlated with prices, while *LSTAT* (lower status percentage) had a negative correlation.

2. Model Performance:

- **Linear Regression:** Provided a basic benchmark with moderate accuracy.

- **Decision Tree Regressor:** Showed improvements over linear regression but had limitations with overfitting.

- **Random Forest Regressor:** Outperformed other models with an RMSE of [Insert RMSE] and R-squared of [Insert R-squared], demonstrating strong performance.

3. Importance of Hyperparameter Tuning:

- Fine-tuning model parameters, especially for Random Forest, significantly enhanced model accuracy and stability.

4. Data Scaling and Encoding:

- Scaling and encoding were essential steps that improved model accuracy, especially for continuous features and categorical variables.

Conclusion

1. Model Selection: The Random Forest model performed best among the evaluated models, achieving the highest accuracy in predicting Boston house prices.

2. Feature Engineering is Key: Creating and selecting relevant features, such as interaction terms and polynomial features, substantially improved predictive power.

3. Data Cleaning and Preprocessing: Proper handling of outliers and normalization of features were crucial in achieving reliable predictions.

4. Hyperparameter Tuning: Tuning played a critical role in refining the model, with Random Forest showing the best results after optimization.

5. Future Directions:

- Explore advanced models like XGBoost or LightGBM, which may offer further improvements.

- Investigate additional features or external datasets to improve accuracy, such as neighborhood crime rates or economic indicators.

Data Summary

1. Train dataset: 506 records with 14 columns, including the target variable `MEDV`.

2. Features: Key features included `RM`, `LSTAT`, `PTRATIO`, `TAX`, and `DIS`.

3. Data Preprocessing: Minimal missing values and some outliers were addressed.

Links and References

- Source Code: GitHub Link

References:

- Scikit-learn Documentation
- Matplotlib Documentation
- Boston Housing dataset