**AIR UNIVERSITY**

**ISLAMABAD**

**Project Report**

**Date 25.12.24**

**Subject: Machine learning**

**Instructor: Dr Masood Anwar**

**Group Members**

Sheikh Usama Iqbal

Tamseen Irfan

Urooj Shoaib

#### **Boston Housing Price Prediction Project Report**

# Objective

The objective of this project is to predict housing prices in the Boston area based on various factors such as the number of rooms, neighborhood characteristics, and socio-economic indicators. Using machine-learning techniques, the goal is to develop models that accurately estimate house prices.

# Define Problem

## Target Variable

The target variable is **MEDV** (Median Value of Owner-Occupied Homes), which represents the median house price in thousands of dollars.

## Goal

To build predictive models that estimate house prices based on available features.

## Business Context

Understanding the factors influencing housing prices provides valuable insights for policymakers, real estate professionals, and urban planners. These insights can help in making data-driven decisions about urban development and economic planning.

# Data Collection

The dataset used is the publicly available **Boston Housing Dataset**, which includes:

* **506 observations** and **14 features** such as crime rate (CRIM), number of rooms (RM), and proximity to employment centers (DIS).
* It contains no missing values, making it ready for analysis and modeling.

# Exploratory Data Analysis (EDA)

During EDA, the following analyses were conducted:

* **Correlation Analysis**:
  + Strong correlation observed between **RM** (number of rooms) and **MEDV**.
  + Negative correlation noted between **LSTAT** (percentage of lower-status population) and **MEDV**.
* **Visualizations**:
  + Plots were used to explore the distribution of MEDV and its relationship with key features like RM and LSTAT.
* **Key Observations**:
  + More rooms are associated with higher house prices.
  + Socio-economic factors like LSTAT significantly impact house prices

# Data Preprocessing

**Handling Missing Values**

* Although the dataset did not contain missing values, preprocessing steps ensured data integrity.

**Feature Engineering**

* Variables with strong correlations to the target variable, such as **RM** and **LSTAT**, were emphasized.
* Interaction terms and polynomial features were explored to capture non-linear relationships.

**Outlier Treatment**

* Outliers in variables like **CRIM** and **TAX** were addressed to improve model stability.

**Scaling**

* Standardization was applied to features, particularly for algorithms sensitive to scale, such as Linear Regression and Gradient Boosting.

**Train-Test Split**

* The dataset was split into:
  + Training Set: 80% of the data
  + Test Set: 20% of the data

# Model Selection

The following machine learning models were applied to predict housing prices:

1. **Linear Regression**
2. **Random Forest Regressor.**
3. **Gradient Boosting Regressor.**

# Model Training

Each model was trained using the training set (80% of the data). During training, model parameters were optimized to best fit the data.

# Model Evaluation

The performance of each model was evaluated on the test set (20% of the data) using various metrics to ensure a comprehensive assessment of their predictive capabilities.

#### **Metrics Used**

* **Mean Squared Error (MSE):** Measures the average squared difference between the predicted and actual values. Lower values indicate better performance.
* **R-Squared (R²):** Represents the proportion of variance in the target variable explained by the model. Higher values indicate better fit.

## Model Results

The following accuracy scores were obtained for each model:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Model** | **MSE** | **R²** | | Linear Regression | 26.01 | 0.645 | | Random Forest Regressor | 8.60 | 0.89 | | Gradient Boosting Regressor | 9.20 | 0.87 | |  |
|  |  |
| **Insights from Evaluation** |  |
| **Random Forest Regressor** is the most reliable model for this dataset, demonstrating a significant improvement over linear methods. Achieved the best results with the lowest MSE (8.60) and highest R² (0.89). Both Random Forest and Gradient Boosting highlight the importance of non-linear modeling for accurate predictions.  **Distribution of Prices(MEDV)** |  |
|  |  |
|  |  |

# Conclusion

This project successfully predicted housing prices in the Boston area using machine-learning techniques. Key findings include:

1. The **Random Forest Regressor** outperformed other models with the lowest MSE (8.60) and highest R² (0.89), making it the most reliable model for predicting house prices.
2. Significant predictors of housing prices include the number of rooms (RM) and the percentage of lower-status population (LSTAT).

The models and insights derived can assist urban planners and real estate professionals in understanding housing price dynamics and making informed decisions.

