CSE 445 Assignment 2 Report

Regression Task with Evaluation Methods

Dataset: Boston Housing (UCI Repository)
Student Name: Adib Ar Rahman Khan
Date: 25/03/25

Objective

The objective of this assignment was to:

- Train and evaluate regression models on the **Boston Housing** dataset,
- Balance both **accuracy** and **interpretability**,
- Explore relationships between variables,
- And apply **improvement techniques** such as scaling, regularization, and non-linear modeling.

Dataset Overview

The dataset includes **506 samples** and **13 features** that describe socioeconomic and environmental attributes of Boston suburbs from 1970. The goal is to predict the **median value of owner-occupied homes (`MEDV`)**.

Data Preprocessing & Exploration

- Loaded the dataset and verified **no missing values**.
- Visualized `MEDV` , which showed a **right-skewed** distribution with a cap at $50,000.
- Used a **correlation heatmap** and **scatter plots** to identify strong predictors:
  - **`RM`** (average rooms) – strong positive correlation
  - **`LSTAT`** (lower-income population %) – strong negative correlation
  - **`NOX`** (pollution) – negative correlation

Modeling & Evaluation

The dataset was split into:

- **70% training**
- **30% testing**

Models Trained & Evaluated:

Model	Type	Scaling	Bonus
Linear Regression	Baseline	No	—
Linear Regression	Re-tested	Yes	Applicable
Polynomial Regression	Non-linear	No	Applicable
Ridge Regression	Regularized (L2)	Yes	Applicable
Lasso Regression	Regularized (L1)	Yes	Applicable
Decision Tree Regressor	Tree-based	No	Applicable
Random Forest Regressor	Ensemble Tree	No	**Best-performing**

Model Performance Summary

Model	MAE	MSE	RMSE	R^2	Adjusted R^2
Linear	3.16	21.52	4.64	0.711	0.684
Linear (Scaled)	3.16	21.52	4.64	0.711	0.684
Polynomial (deg=2)	3.06	25.26	5.03	0.661	0.629
Lasso	3.21	22.79	4.77	0.694	0.665
Ridge	3.16	21.55	4.64	0.711	0.684
Decision Tree	2.46	10.83	3.29	0.855	0.841
Random Forest	**2.10**	**9.71**	**3.12**	**0.870**	**0.857**

Feature Importance

- **Linear Regression**: `RM`, `CHAS`, `NOX`, and `LSTAT` were the strongest drivers of `MEDV`.
- **Random Forest**: Identified `LSTAT`, `RM`, and `DIS` as highly important through ensemble splitting.

Visualizations of feature importance were provided using **bar plots**, highlighting both the direction and strength of influence.

Bonus Improvements Attempted

Technique	Result
Feature Scaling	No effect on linear models
Polynomial Regression	Increased complexity, but worse R^2
Ridge & Lasso	Slight regularization, no gain
Decision Tree	Strong non-linear capture
Random Forest	Most accurate & robust

Conclusion:

- **Random Forest Regressor** is the **best-performing model**, excelling in both error reduction and variance explanation.
- **Linear Regression** offers speed and interpretability but fails to capture complex patterns.
- **Polynomial Regression** adds complexity without improving accuracy.
- Regularization has minimal impact, suggesting the dataset is not suffering from overfitting.