Boston Housing Price Prediction - Project Documentation

1. Project Overview

This project focuses on predicting house prices in the Boston area using machine learning techniques.

The goal is to build regression models to estimate housing prices and analyze key influencing factors.

2. Objectives

- Understand relationships between features and prices.
- Build Linear, Ridge, and Polynomial Regression models.
- Evaluate performance using MSE and R^2.
- Interpret models using SHAP explainability.

3. Technologies Used

Python, pandas, numpy, scikit-learn, seaborn, matplotlib, shap

4. Dataset Description

Boston Housing dataset (OpenML)

Target variable: PRICE

Key features: CRIM, RM, TAX, PTRATIO, LSTAT

5. Project Structure

data/ - dataset

notebooks/ - Jupyter notebooks for models

visuals/ - plots and figures

models/ - saved trained models

reports/ - project summaries

6. Methodology

- Data preprocessing: cleaning, scaling, outlier removal
- Feature engineering: TAX/RM, AGE/RM, etc.

- Model building: Linear, Ridge, Polynomial Regression
- Evaluation: MSE, R^2
- Explainability: SHAP visualization

7. Results Summary

Ridge Regression achieved the best R^2 (~0.82).

Linear Regression baseline: R^2 ~0.70

Polynomial Regression: R^2 ~0.79

8. Key Visualizations

- Correlation Heatmap
- Actual vs Predicted Prices
- SHAP Beeswarm Plot

9. Outcomes

RM (number of rooms) and LSTAT (lower status %) are strong predictors.

Ridge Regression reduced overfitting and improved accuracy.

10. Future Enhancements

- Add Lasso or ElasticNet models
- Deploy with Streamlit/Flask
- Try tree-based models like Random Forest

11. References

- Scikit-learn Documentation
- OpenML Boston Dataset
- SHAP Library