CS6220 - Data Mining Techniques

Project Report

Project Title: Predicting Boston Housing Prices using Regression Analysis

Team : Lemon Grass

By:

Deep Rahul Shah

Orijeet Mukherjee

Pravin Anand Pawar

Preksha Patil

Abstract

This project aims to predict housing prices in the Boston area by analyzing historical data and understanding the impact of factors such as crime rate, number of rooms, and socioeconomic status on median home values. Through exploratory data analysis, feature importance analysis, and the application of multiple machine learning models, this project seeks to identify the best-performing predictor of housing prices. Ethical considerations regarding the use of sensitive data such as socioeconomic status are also explored, ensuring the analysis aligns with responsible AI principles.

Introduction

The prediction of housing prices is a critical aspect of real estate economics, significantly influencing decisions by buyers, investors, and policymakers. This project focuses on predicting Boston housing prices using a historical dataset containing socioeconomic and housing characteristics. Factors such as crime rates, proximity to employment centers, and number of rooms provide key insights into the determinants of housing prices. By applying machine learning, this project aims to uncover influential features, optimize predictive models, and understand housing market dynamics.

Problem Statement:

The main goal of this project is to accurately predict the median housing prices in Boston suburbs. Specifically, it seeks to:

1. Analyze how factors such as crime rate, room numbers, and socioeconomic conditions influence housing prices.
2. Develop and evaluate machine learning models for housing price prediction.
3. Address ethical concerns related to the inclusion of socioeconomic features in predictive modeling.

The outcomes are intended to provide actionable insights into housing affordability and market trends while ensuring responsible use of data.

Motivation

Understanding the factors influencing housing prices helps promote market efficiency, affordability, and informed decision-making. This project combines data mining techniques with machine learning models to uncover these drivers.

Background and Literature Review

Predicting housing prices has been a widely studied problem in real estate economics. Past studies using the Boston Housing dataset have demonstrated the importance of variables like the average number of rooms, crime rates, and accessibility to employment centers. Machine learning models such as Random

Forest and Neural Networks have been particularly effective in capturing complex, nonlinear relationships.

Goal: The primary goal is to compare multiple machine learning models and identify the best-performing algorithm for predicting housing prices. Additionally, the project aims to address ethical considerations, such as the potential biases in socioeconomic data.

Data Source: The Boston Housing dataset is used, which contains information about various attributes of houses in Boston suburbs, including crime rate, average number of rooms, and distance to employment centers. This dataset helps in predicting the median value of homes and understanding the key factors that affect these values.

Dataset Description: The Boston Housing dataset contains 506 entries and 14 features, including:

1. Target variable: medv (median value of homes in $1000s).
2. Key features:
    - rm: Average number of rooms per dwelling.
    - lstat: Percentage of lower-status population.
    - crim: Crime rate per capita.

The dataset is widely used for regression tasks in academic and machine learning contexts.

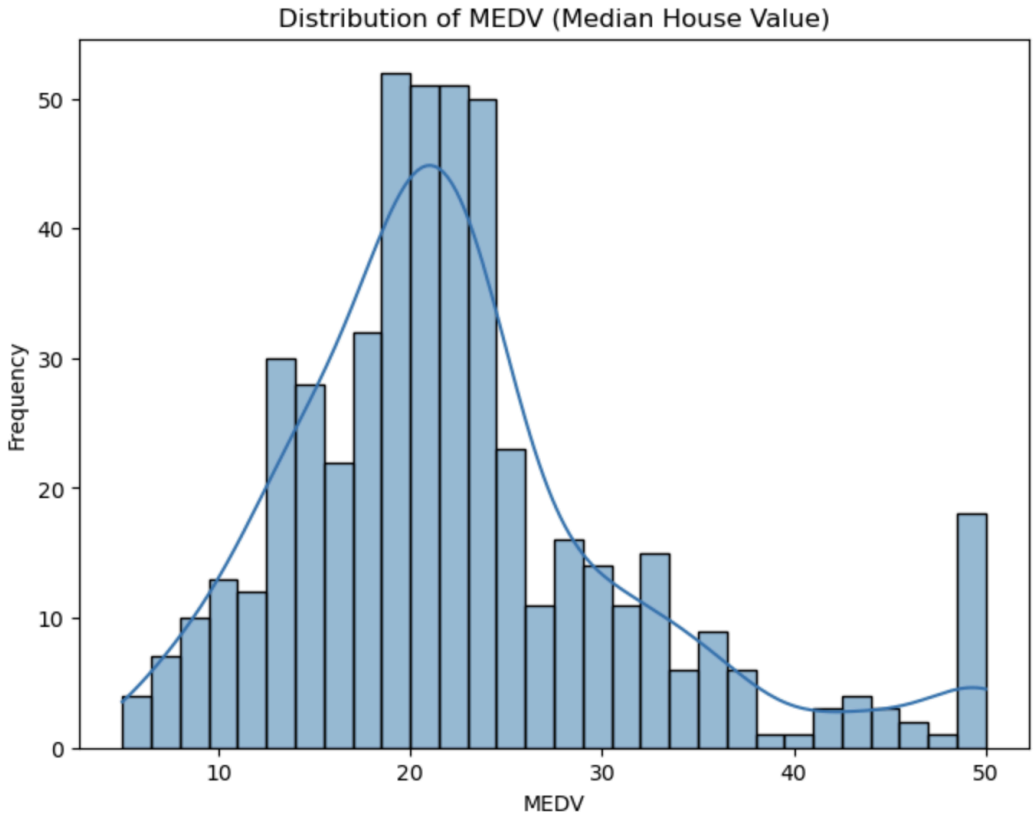
Methodology

1. Exploratory Data Analysis (EDA)

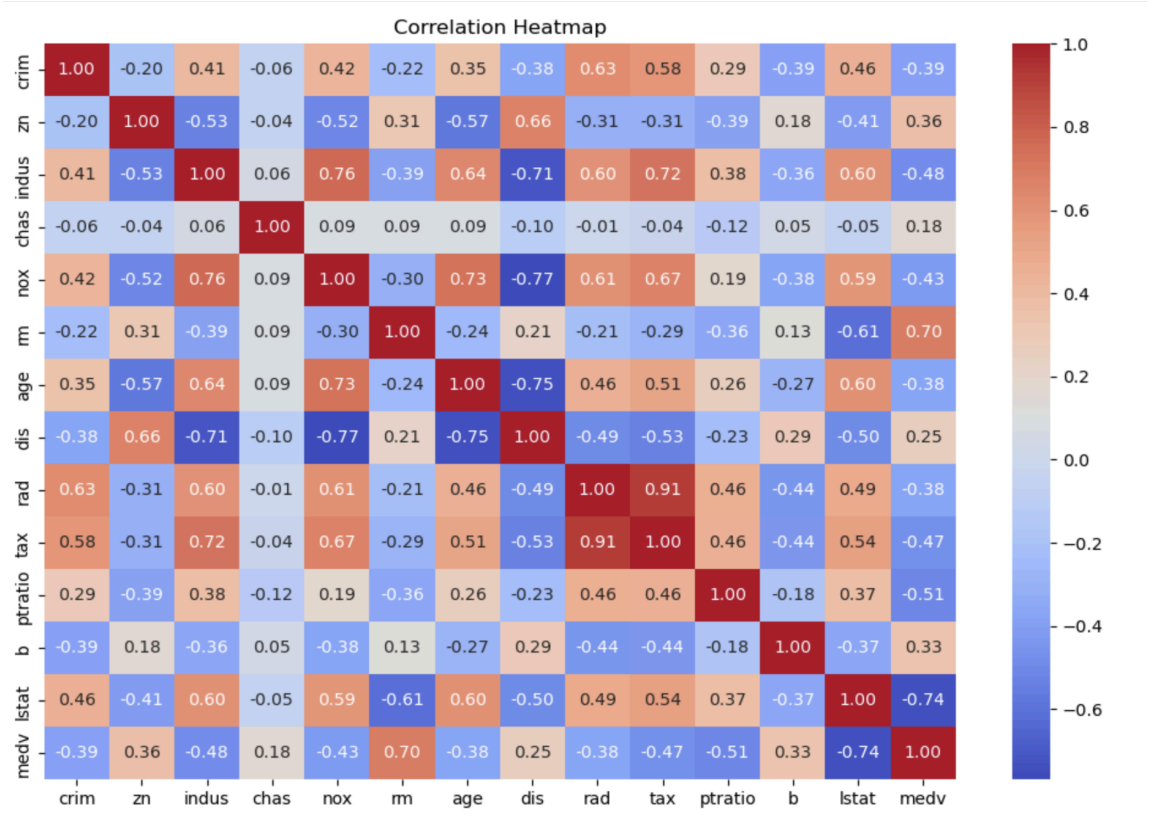
Objective: To analyze distributions, relationships, and detect outliers.

Insights:

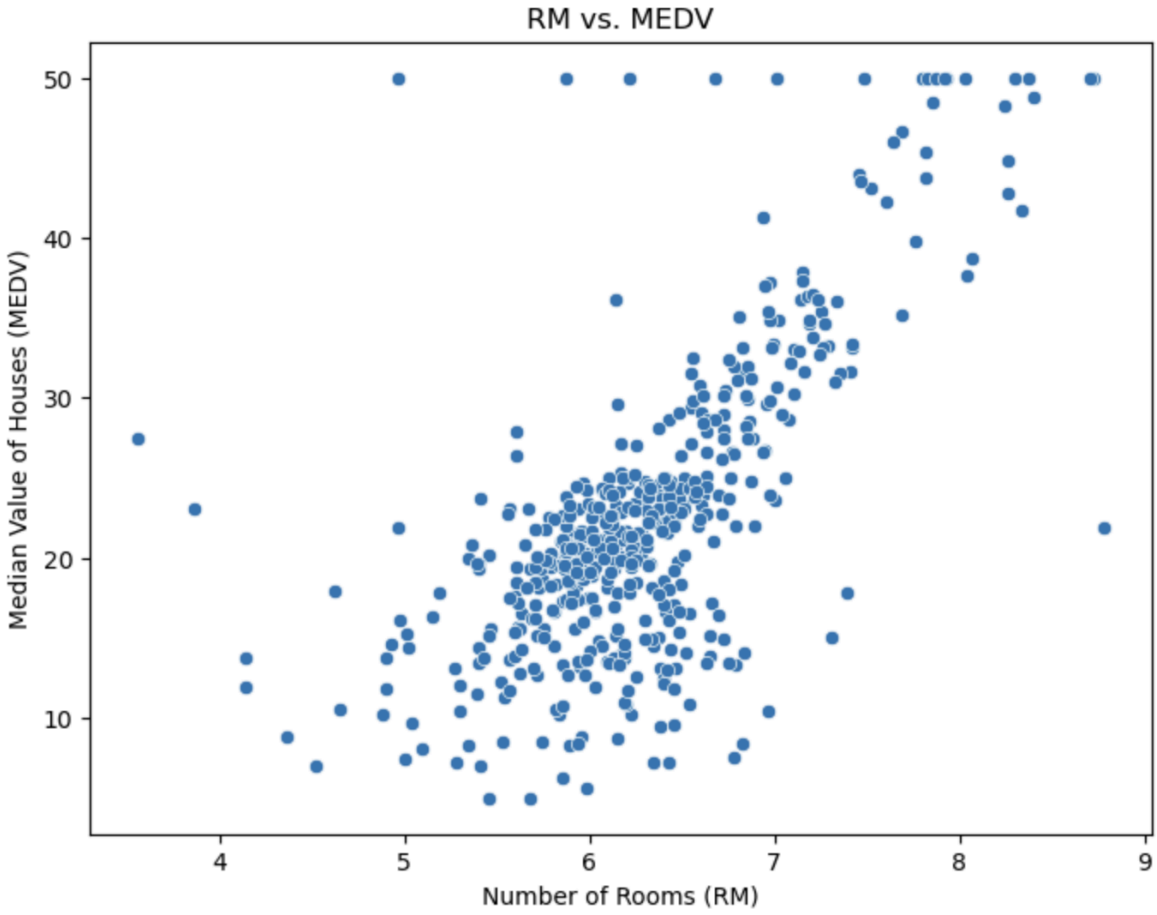
- rm has a strong positive correlation with medv, indicating that homes with more rooms are more expensive.
- lstat and crim have negative correlations with medv, suggesting that higher crime rates and lower socioeconomic status reduce housing prices.
- Target variable (medv): Represents the median house value, with a positively skewed distribution.
- Outliers in crim and tax were detected and addressed during preprocessing.
- Distribution of medv (median house value) showing a positively skewed distribution.



- Correlation heatmap showing relationships between features and medv. Features like rm show a strong positive correlation, while lstat and crim exhibit strong negative correlations.



- Scatterplot showing the positive relationship between rm (average number of rooms) and medv (median house value).



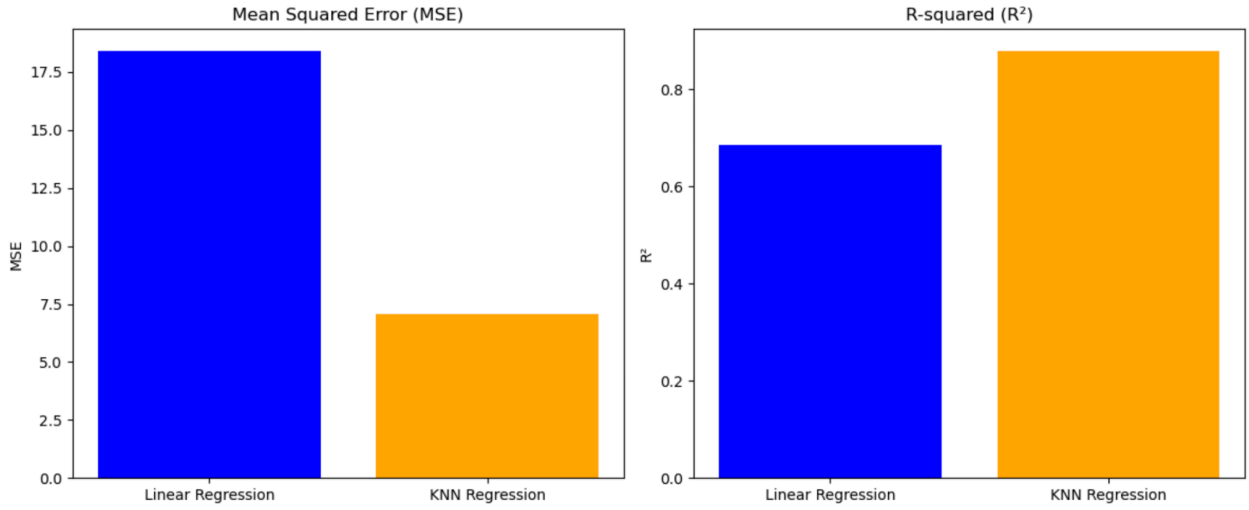
2. Feature Engineering

- **Feature Importance Analysis:** Using SHAP to determine the contributions of individual features.
- **Dimensionality Reduction:** Principal Component Analysis (PCA) reduced the dataset to 10 principal components while retaining 95% of the variance.

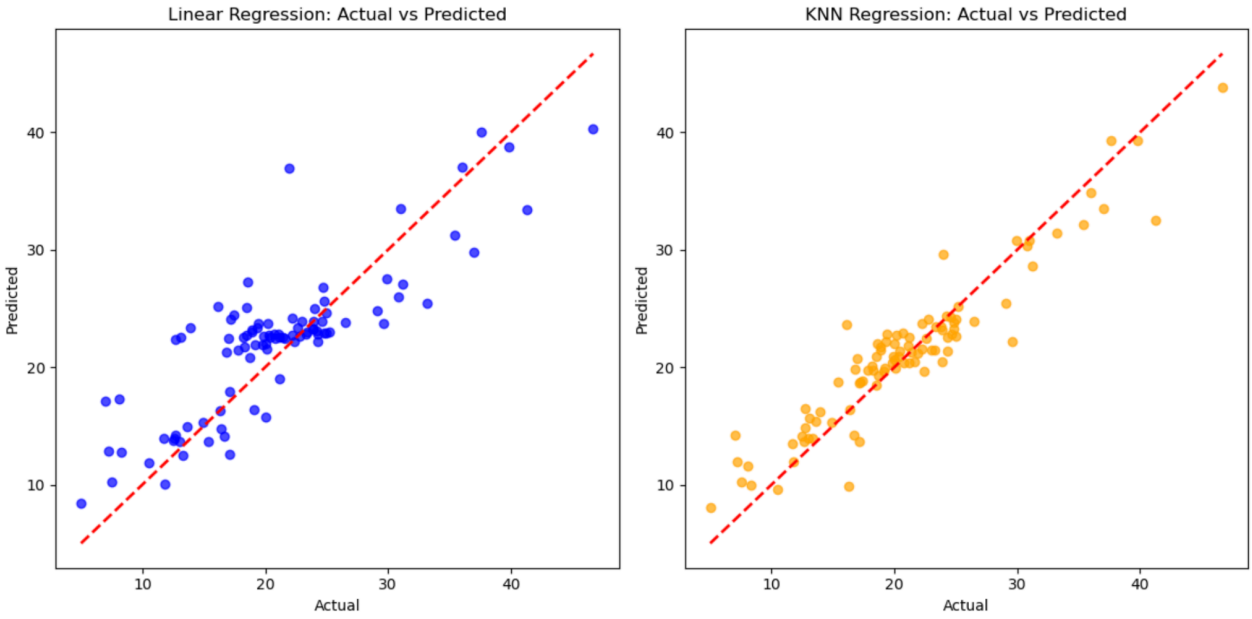
3. Model Development

- Develop four different Models:
    - **Linear Regression**: Simple, interpretable, limited for non-linear data.
    - **KNN**: Predicts using nearest neighbors; sensitive to scaling and k.
    - **Random Forest**: Ensemble trees; robust, handles non-linearity.
    - **Neural Networks**: Deep model; captures complex patterns; requires tuning.
- Best-performing model selected based on these metrics: highest R^2, lowest RMSE/MAE.
- SHAP plots to explain the contribution of each feature to a machine learning model's prediction, enhancing model interpretability and transparency.
- Model performance comparison: Mean Squared Error (MSE) and R-squared (R^2) for Linear Regression and KNN. KNN outperformed Linear Regression with significantly lower MSE and higher R^2.

```
Model Comparison:
              Model  Mean Squared Error (MSE)  R-squared (R²)
0  Linear Regression                 18.422169        0.684924
1     KNN Regression                  7.061042        0.879234
```



- Scatterplots showing Actual vs. Predicted values for Linear Regression and KNN. KNN predictions align more closely with the diagonal line, indicating better performance.



Code

The implementation includes:

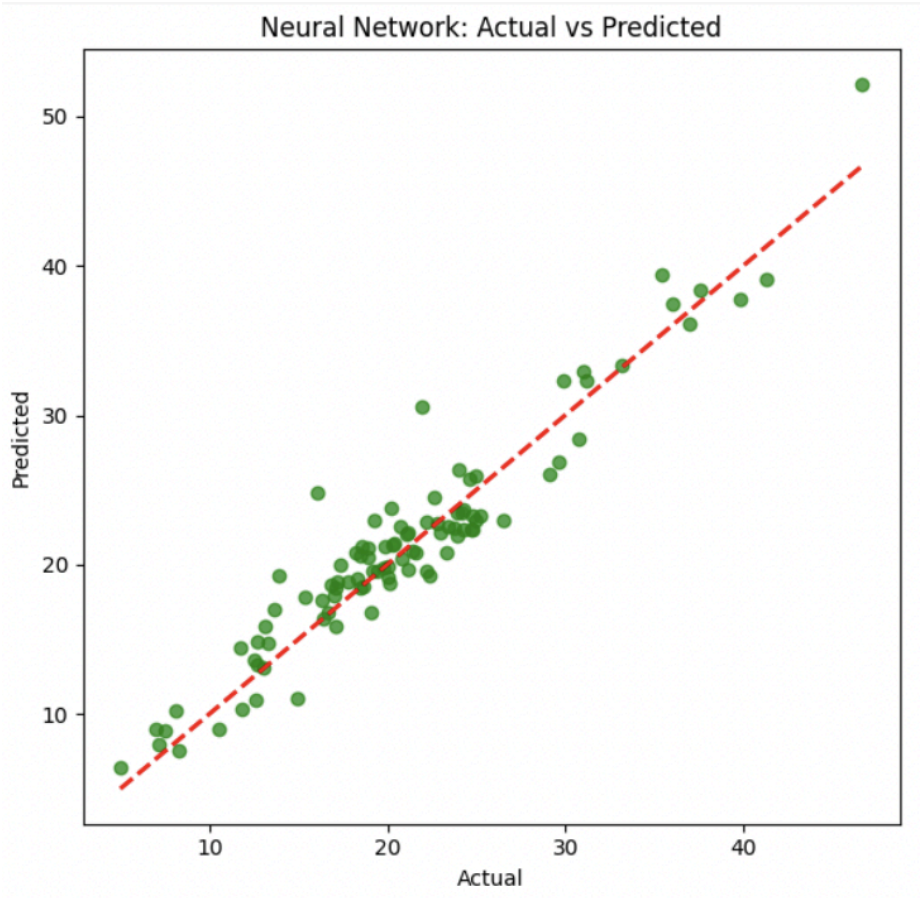
1. Preprocessing Pipeline: Missing value imputation, feature scaling, polynomial feature generation, and outlier detection.
2. Dimensionality Reduction: PCA for efficient feature representation.
3. Model Development: Training Linear Regression, K-Nearest Neighbors, Random Forest, and Neural Network models.
4. Evaluation: Metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2.

Results:

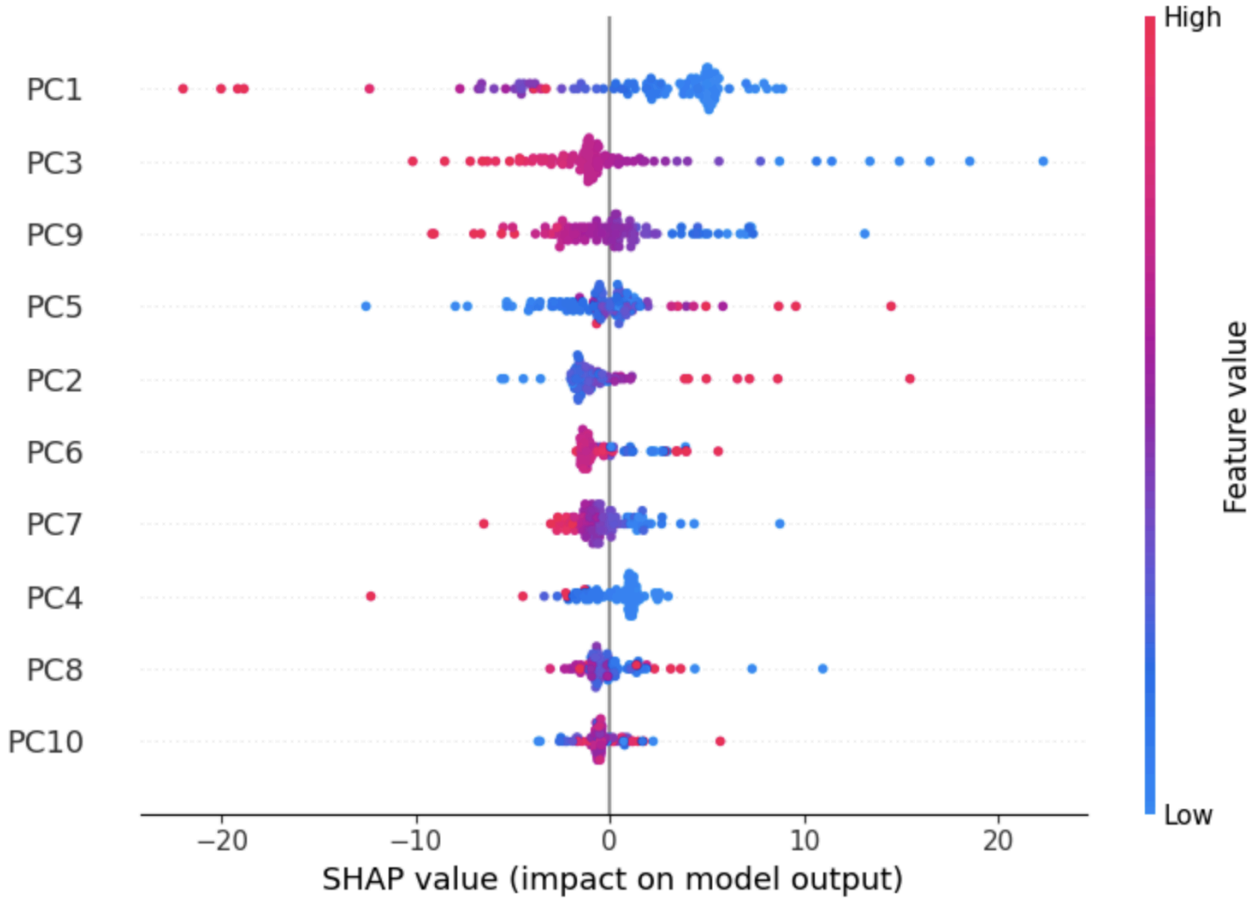
- Neural Network model was found to be the best model with highest R^2 value of 0.90 and lowest RMSE value of 2.33, followed by KNN for the second best model. From SHAP plot, we can see PC3 and PC9 components has shown strong potential for this regression task of price prediction. Thus, Neural Network has shown strong potential to deploy for real world house price prediction applications.
- **Model Performance:**

```
Linear Regression: RMSE = 4.2921, MSE = 18.4222, R² = 0.6849
KNN Regression: RMSE = 2.6573, MSE = 7.0610, R² = 0.8792
Neural Network: RMSE = 2.3344, MSE = 5.4496, R² = 0.9068
Random Forest: RMSE = 4.5672, MSE = 20.8596, R² = 0.6432
```

- **Neural Network Prediction:** prediction values close to true values.



- **Best model SHAP plot:**



Discussion:

- **Strengths:**
  - Found Neural Network as optimal model for Boston house price prediction.
  - Insights into feature significance.
- **Limitations:**
  - Small dataset size.
  - Potential overfitting in tree-based models.

Conclusion:

• Successfully built predictive models for Boston housing prices.
• Highlighted key features affecting housing values.
• Neural Network is the optimal model for this regression task.
• Showed potential for applying similar techniques in real-world scenarios.

Future Work:

Using large and diverse real-world dataset for Boston house price modelling.

References:

- National Association of Realtors. (2024). Housing affordability challenges and urban price trends. Retrieved from https://www.nar.realtor
- Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
- Housing Data Research. (2023). Impact of socio-economic factors on real estate pricing. Journal of Urban Economics, 19(2), 112-130.
- Kaggle. (2024). Boston Housing Dataset Analysis. Retrieved from https://www.kaggle.com

Expected Deliverables:

- Complete project documentation, including source code, detailed report (with analysis, methodology, experiments, plots, etc.), PowerPoint presentation, and a recorded video presentation.
- A comprehensive analysis of the factors affecting housing prices.
- All four fine-tuned ML models are each capable of accurately predicting housing prices, and comparative analysis to identify the best-suited ML algorithm for predicting housing prices.
- Insights into the ethical implications of using socioeconomic data in predictive modeling.