By group “Lemon Grass”

CS6220 - Data Mining Techniques

Predicting Boston Housing Prices using Regression Analysis

8th December 2024

Team Members

Deep Rahul Shah

Orijeet Mukherjee

Pravin Anand Pawar

Preksha Patil

Introduction

- Housing prices impact urban planning, policy-making, and economic stability.
- According to the National Association of Realtors, housing affordability is a growing concern, with a 20% year-over-year price increase in some urban areas.
- Accurate predictions can guide better decision-making for buyers, sellers, and policymakers.

Problem Statement

Real estate pricing is influenced by complex, nonlinear factors.

Questions:

1. How accurately can we predict housing prices using regression models?
2. Which features have the most significant impact on price predictions?

Goal: Build a model with high predictive accuracy and interpretability.

Data Source

Dataset: Boston Housing (506 samples, 13 features)

Key Features:

1. CRIM (crime rate per capita)
2. RM (average number of rooms)
3. LSTAT (% lower status population)

Target: MEDV (Median value of owner-occupied homes)

Methodology

Exploratory Data Analysis (EDA)

- Analyze distributions, relationships, and outliers.
- Visualize data insights.

Feature Engineering:

- Feature importance analysis (e.g., SHAP).
- Dimensionality reduction with PCA.

Model Development:

- Algorithms used: Linear Regression, KNN, Random Forest, Neural Networks.
- Evaluation metrics: Mean Absolute Error, Root Mean Square Error, R-Squared

Data Pre-Processing

Loading the Dataset and Handling Missing Values

- Initial checks for missing values, data types, and basic structure
- Missing Values Strategy: Replace missing values with the mean using SimpleImputer.

Feature Scaling:

- Technique: Standardization using StandardScaler to normalize data.

Feature Engineering:

- Created polynomial features up to degree 2 for interaction terms.

Outlier Detection and Removal

- Method: IsolationForest with 5% contamination threshold.
- Result: Removed outlier rows, ensuring cleaner data for modeling.

Dimensionality Reduction

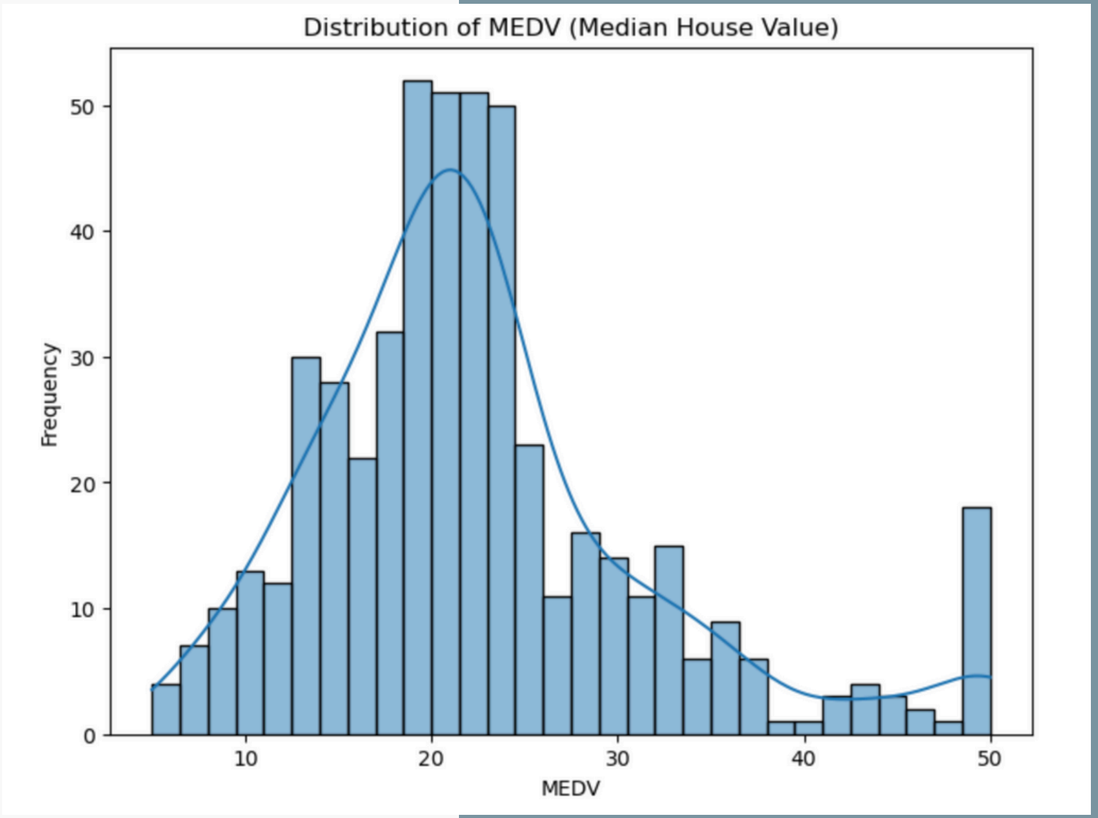
- Applied PCA to reduce dimensions to 10 principal components while retaining variance.
- Partitioned the dataset into an 80:20 training-testing split.

Exploratory Data Analysis (EDA)

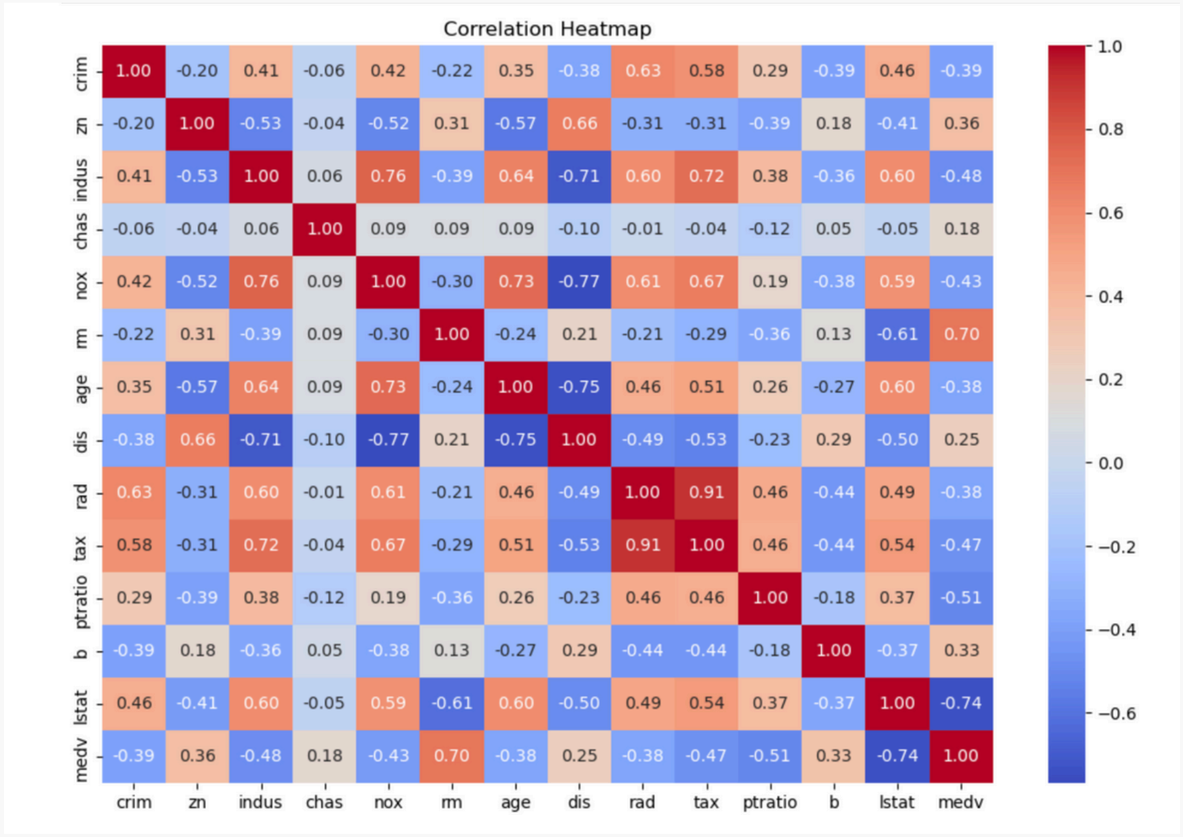
- The dataset contains 506 entries with 14 features, including socioeconomic and housing-related variables.
- Target variable (medv): Represents the median house value, with a positively skewed distribution.
- Strong feature correlations:

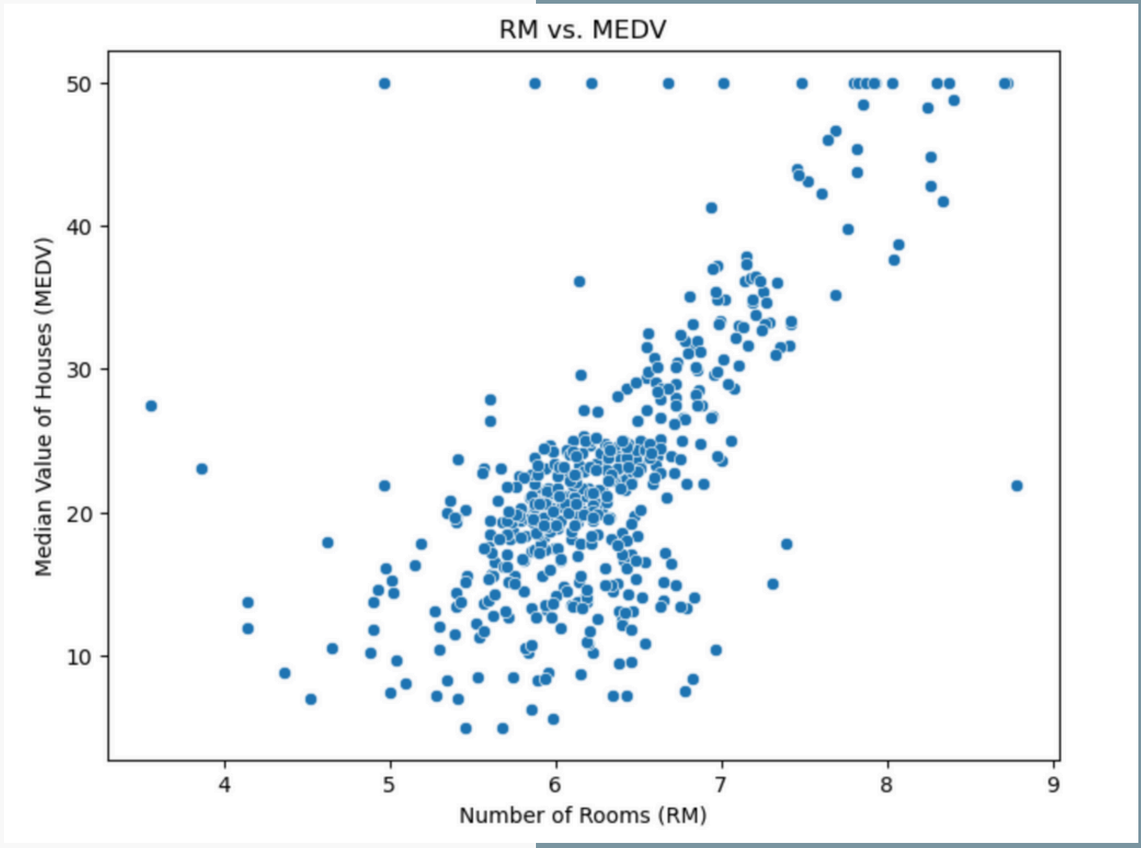
1. rm (average number of rooms) has a strong positive correlation with medv.
2. lstat (percentage of lower status population) has a negative correlation with medv.

- Outliers detected in features like crim (crime rate) and tax (property tax).



Exploratory Data Analysis (EDA)





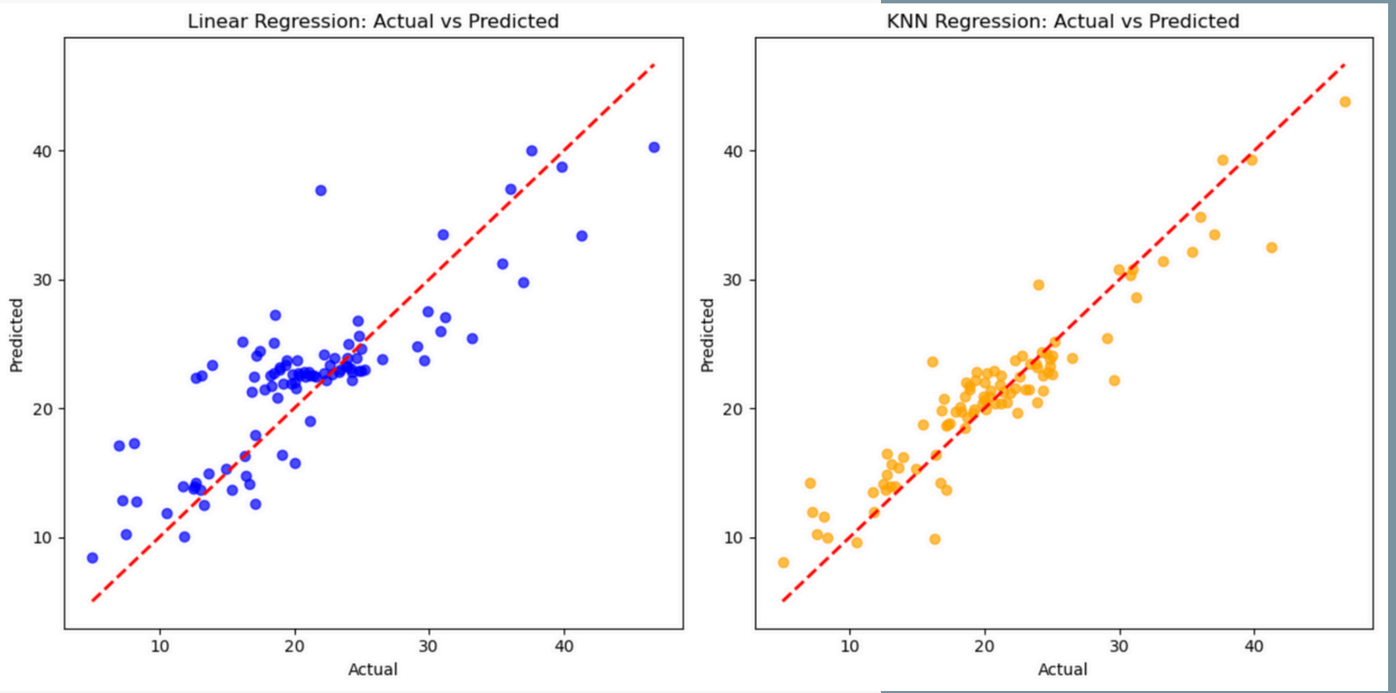
Model Performance: Linear Regression and KNN

Key Metrics:

- Linear Regression:
- Mean Squared Error (MSE): 18.42
- R-squared (R^2): 0.68

KNN Regression:

- Mean Squared Error (MSE): 7.06
- R-squared (R^2): 0.88



Bar Chart Comparison:

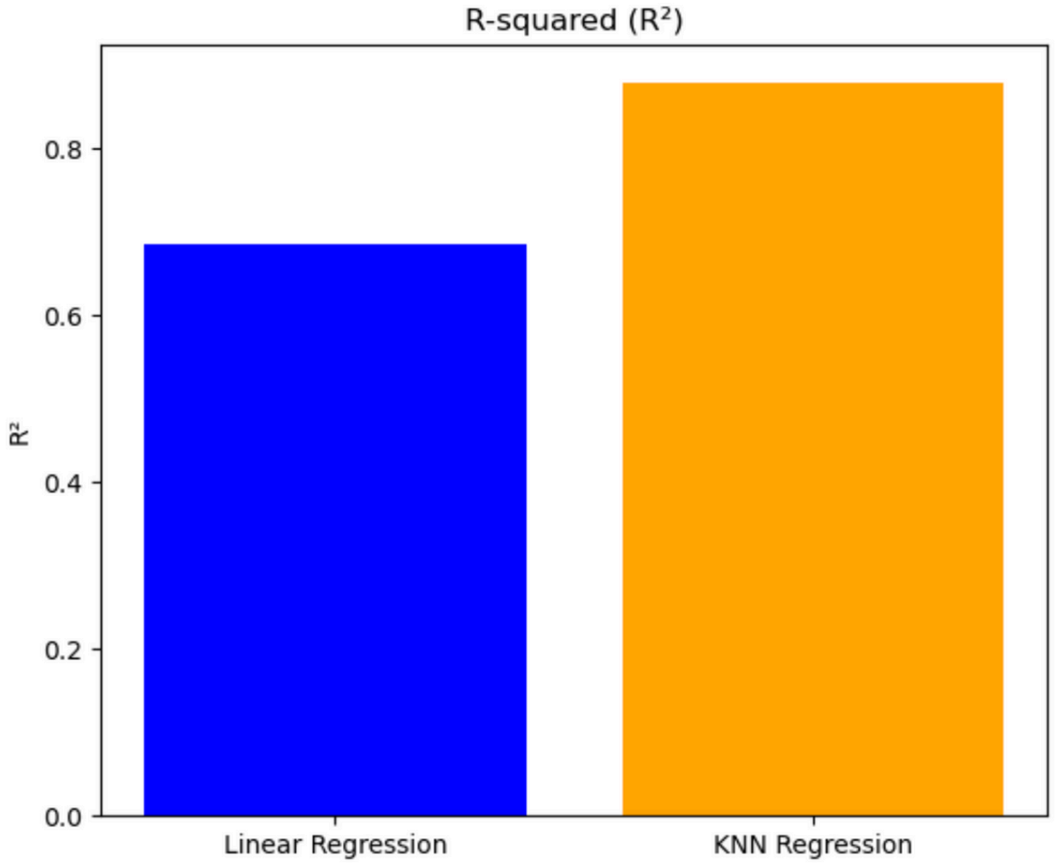
- Displays MSE and R2R^2R2 for both models, highlighting KNN's better performance.

Scatterplots:

- Linear Regression: Actual vs. Predicted values show moderate prediction alignment.
- KNN Regression: Actual vs. Predicted values align closely along the diagonal.

Model Comparison:

	Model	Mean Squared Error (MSE)	R-squared (R^2)
0	Linear Regression	18.422169	0.684924
1	KNN Regression	7.061042	0.879234



Model Performance: Random Forest and Neural Networks

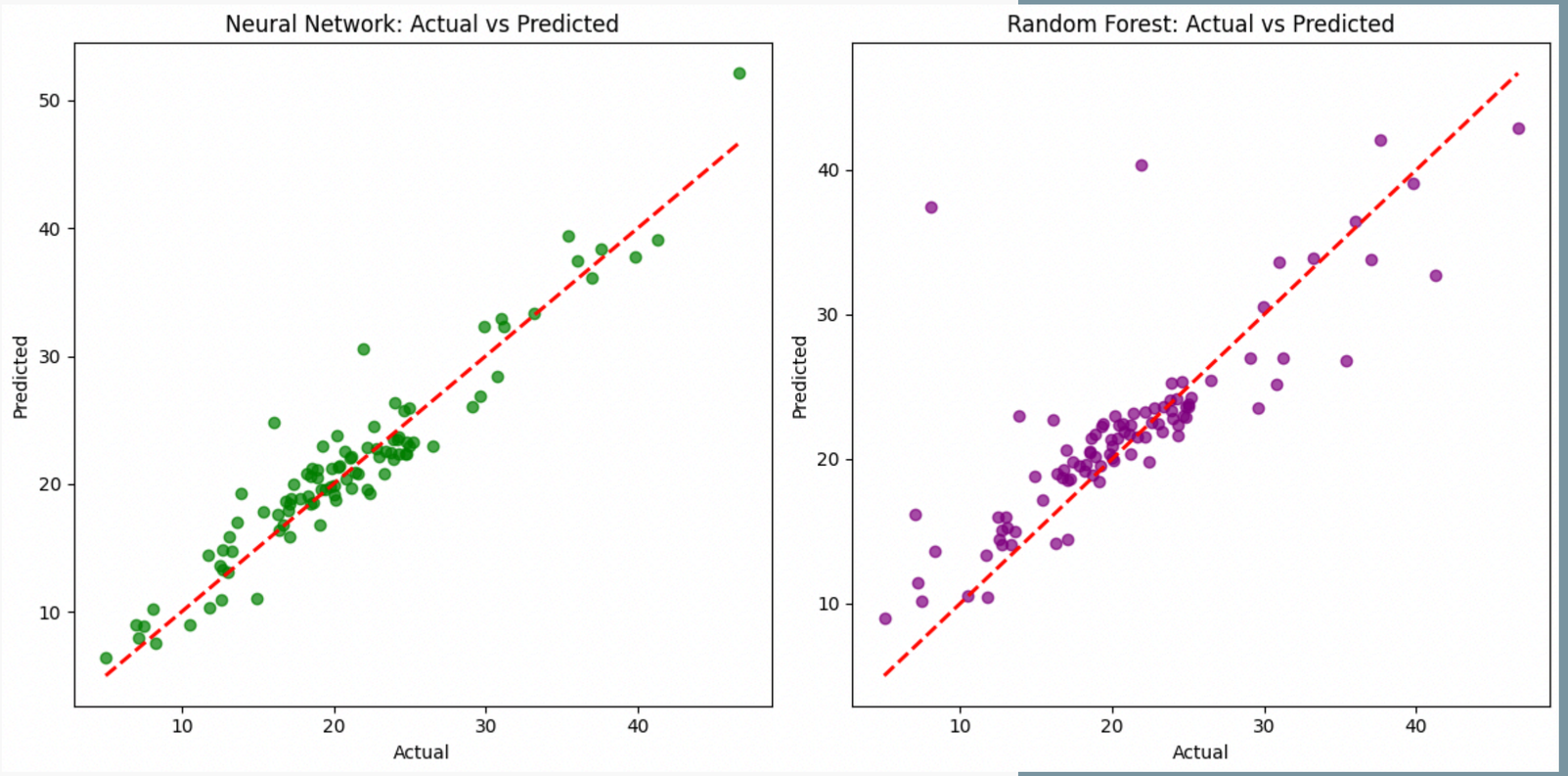
Key Metrics:

Random Forest:

- Mean Squared Error (MSE): 20.85
- R-squared (R^2): 0.64

Neural Networks:

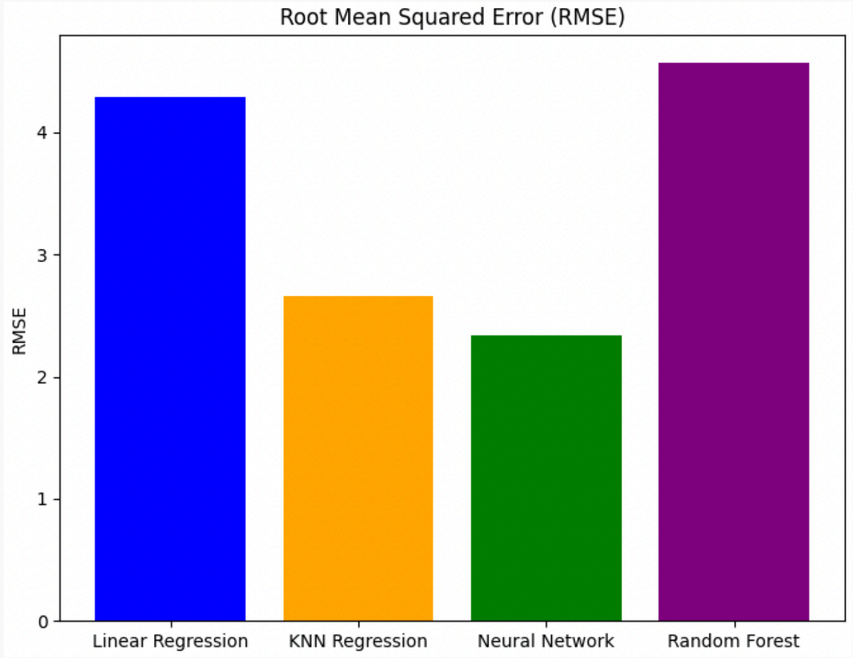
- Mean Squared Error (MSE): 5.44
- R-squared (R^2): 0.90



Scatterplots:

- Random Forest:
  - Actual vs. Predicted values show moderate prediction performance with some predictions far from true value
- Neural Network:
  - Actual vs. Predicted values are move close to true values.

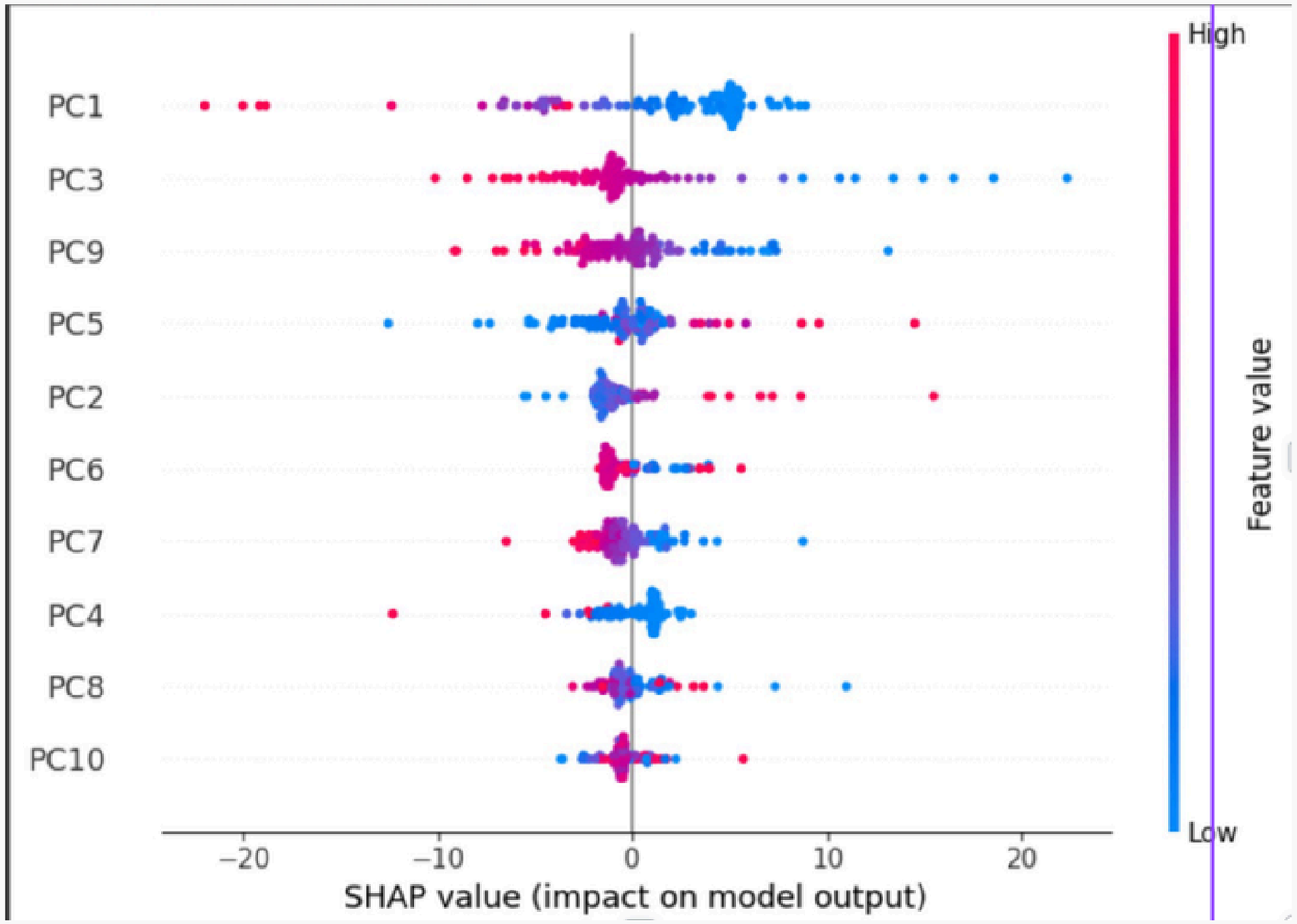
Model Performance Comparison




```
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 Networks is best performing model based on R^2 values.

SHAP plot


PC1
PC3
PC9
PC5
PC2
PC6
PC7
PC4
PC8
PC10
−20
−10
0
10
20
SHAP value (impact on model output)
High
Low
Feature value

Discussions

Strengths:

- Best R^2 value found for Neural Network model.
- Insights into feature significance.

Limitations:

- Small dataset size.
- Potential overfitting in tree-based models.

Future Work:

- Using larger, more diverse datasets.

Conclusion

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

References and Acknowledgement

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

Acknowledgement:

- Prof. Sara Arunangiri
- Teaching Assistants

Thank you