REPORT

SUMMARIZING THE DATA AND ITS PROCESSING

1. The Dataset and Its Features:

CRIM: Crime rate per capita by town.

ZN: Percentage of residential land zoned for lots over 25,000 sq.ft.

INDUS: Percentage of non-retail business acres per town.

CHAS: Charles River dummy variable (1 if tract bounds river; 0 otherwise).

NOX: Nitric oxides concentration (parts per 10 million).

RM: Average number of rooms per dwelling.

AGE: Percentage of owner-occupied units built before 1940.

DIS: Weighted distances to five Boston employment centers.

RAD: Index of accessibility to radial highways.

TAX: Full-value property-tax rate per \$10,000.

PTRATIO: Pupil-teacher ratio by town.

B: $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town.

LSTAT: Percentage of lower status of the population.

MEDV: Median value of owner-occupied homes in \$1000's. (Output)

2. Data Preprocessing Steps:

Data Type Conversion: Changed data types from object to float for analysis

and operations.

Handling Outliers: Identified and capped outliers in relevant columns.

Constant Column Check: Checked for and removed any constant columns.

Correlation Analysis: Analyzed and removed highly correlated columns to

avoid multicollinearity.

Skewness Check: Assessed skewness in the data.

Data Transformation: Transformed columns to improve model performance.

Train-Test Split: Split the dataset into training and test sets.

Data Scaling: Applied Standard Scaler for scaling the data.

3. Model Training and Evaluation Results:

Linear Regression: Achieved an R² score of 81%.

XG Boost: Recorded an adjusted R² score of 82.08%.

Decision Tree Regressor: Recorded an adjusted R² score of 74.07%.

Gradient Boosting Regressor: Recorded an adjusted R² score of 84.11%.

Random Forest Regressor: Delivered an adjusted R² score of 84.17%.

Hyper-Parameter Tuning: Grid Search CV and Randomized Search CV on Random Forest resulted in an adjusted R² score of 82.08%, which was lower

than the default Random Forest model. So it was not taken.

4. Interpretation of Model Performance and Coefficients:

- Concluded that the Random Forest model is the best, with an adjusted R^2 score of 84.17% .
- Identified important features using Random Forest and noted that some features contributed minimally to the output.
- After removing insignificant features, the adjusted R² score improved to 84.41%.

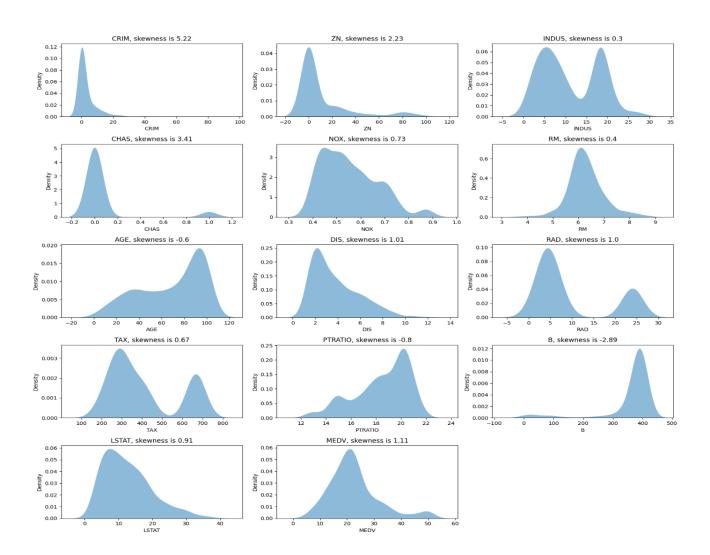
5. Challenges Faced:

The primary challenge was that hyper-parameter tuning degraded the model's performance.

6. visualised results

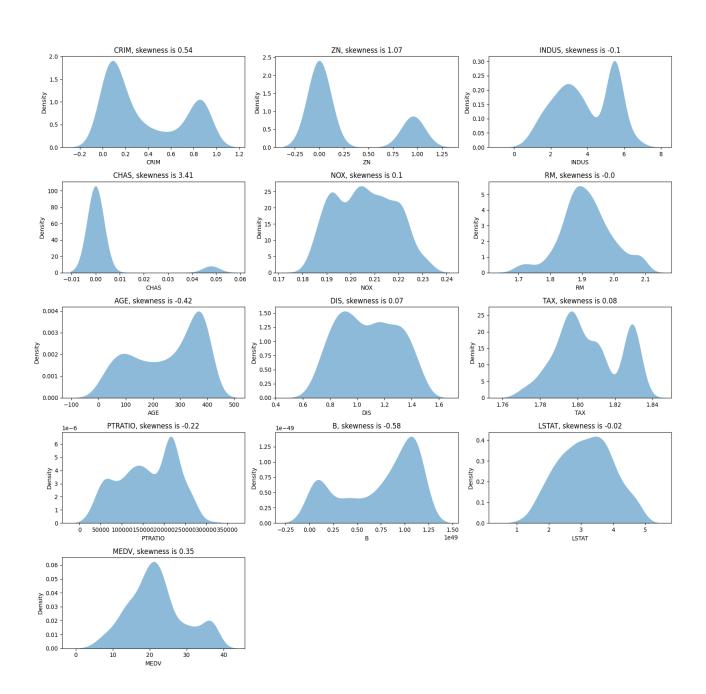
Distribution before skewness removal

Boston House Prices: Distribution Analysis



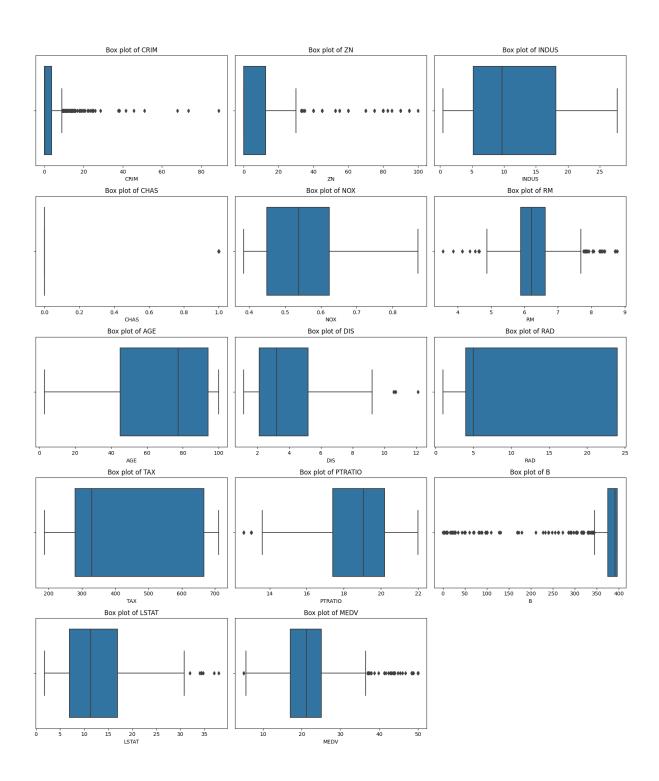
Distribution after skewness removal

Boston House Prices: Distribution Analysis After Skewness Treatment



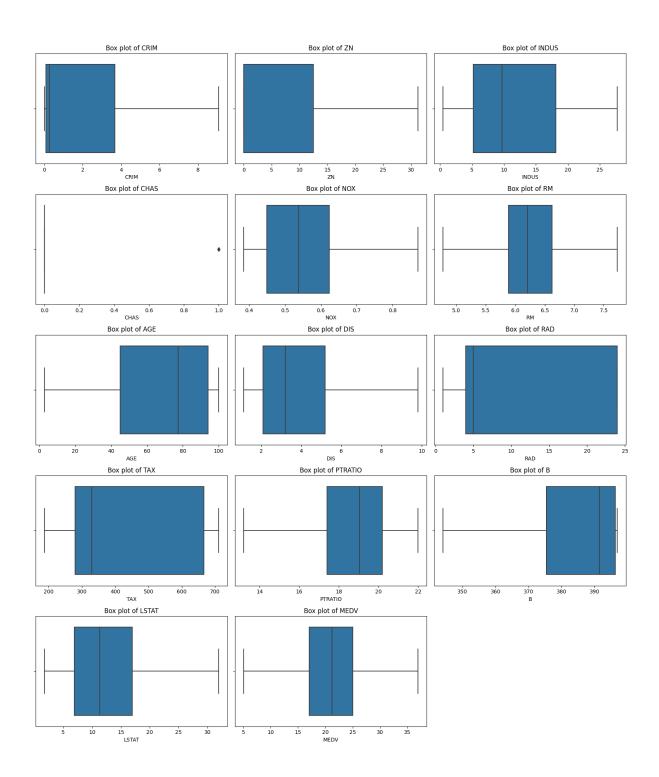
Boxplot before Outlier Removal

Box Plots of Each Column



Boxplot after Outlier Removal

Box Plots of Each Column



HeatMap before corelated feature removal

CRIM	1	-0.37	0.63	-0.031	0.66	-0.26	0.52	-0.56	0.93	0.87	0.42	-0.44	0.59	-0.54
NZ -	-0.37	1	-0.58	-0.038	-0.55	0.36	-0.58	0.68	-0.34	-0.38	-0.43	0.24	-0.45	0.43
SUDNI	0.63	-0.58	1	0.063	0.76	-0.41	0.64	-0.71	0.6	0.72	0.38	-0.42	0.61	-0.56
CHAS	-0.031	-0.038	0.063	1	0.091	0.082	0.087	-0.099	-0.0074	-0.036	-0.12	-0.011	-0.053	0.15
XOX -	0.66	-0.55	0.76	0.091	1	-0.32	0.73	-0.78	0.61	0.67	0.19	-0.43	0.6	-0.51
R -	-0.26	0.36	-0.41	0.082	-0.32	1	-0.26	0.22	-0.21	-0.29	-0.35	0.2	-0.63	0.7
AGE	0.52	-0.58	0.64	0.087	0.73	-0.26	1	-0.75	0.46	0.51	0.26	-0.31	0.61	-0.46
DIS	-0.56	0.68	-0.71	-0.099	-0.78	0.22	-0.75	1	-0.5	-0.54	-0.24	0.31	-0.5	0.33
RAD -	0.93	-0.34	0.6	-0.0074	0.61	-0.21	0.46	-0.5	1	0.91	0.47	-0.39	0.49	-0.45
TAX	0.87	-0.38	0.72	-0.036	0.67	-0.29	0.51	-0.54	0.91	1	0.46	-0.43	0.55	-0.54
PTRATIO	0.42	-0.43	0.38	-0.12	0.19	-0.35	0.26	-0.24	0.47	0.46	1	-0.1	0.38	-0.52
а -	-0.44	0.24	-0.42	-0.011	-0.43	0.2	-0.31	0.31	-0.39	-0.43	-0.1	1	-0.35	0.32
LSTAT	0.59	-0.45	0.61	-0.053	0.6	-0.63	0.61	-0.5	0.49	0.55	0.38	-0.35	1	-0.8
MEDV	-0.54	0.43	-0.56	0.15	-0.51	0.7	-0.46	0.33	-0.45	-0.54	-0.52	0.32	-0.8	1
	CRIM	ΖN	INDUS	CHAS	NOX	RM	AĞE	DİS	RAD	TAX	PTRATIO	В	LSTAT	MEDV

- 0.0

- -0.6

HeatMap after corelated feature removal

														- 1	
CRIM	1	-0.37	0.63	-0.031	0.66	-0.26	0.52	-0.56	0.87	0.42	-0.44	0.59	-0.54		.0
NZ -	-0.37	1	-0.58	-0.038	-0.55	0.36	-0.58	0.68	-0.38	-0.43	0.24	-0.45	0.43	- 0	.8
INDUS	0.63	-0.58	1	0.063	0.76	-0.41	0.64	-0.71	0.72	0.38	-0.42	0.61	-0.56	- 0).6
CHAS	-0.031	-0.038	0.063	1	0.091	0.082	0.087	-0.099	-0.036	-0.12	-0.011	-0.053	0.15		
NOX -	0.66	-0.55	0.76	0.091	1	-0.32	0.73	-0.78	0.67	0.19	-0.43	0.6	-0.51	- 0	.4
M -	-0.26	0.36	-0.41	0.082	-0.32	1	-0.26	0.22	-0.29	-0.35	0.2	-0.63	0.7	- 0).2
AGE	0.52	-0.58	0.64	0.087	0.73	-0.26	1	-0.75	0.51	0.26	-0.31	0.61	-0.46		
DIS -	-0.56	0.68	-0.71	-0.099	-0.78	0.22	-0.75	1	-0.54	-0.24	0.31	-0.5	0.33	- 0	.0
TAX	0.87	-0.38	0.72	-0.036	0.67	-0.29	0.51	-0.54	1	0.46	-0.43	0.55	-0.54		-0.2
PTRATIO	0.42	-0.43	0.38	-0.12	0.19	-0.35	0.26	-0.24	0.46	1	-0.1	0.38	-0.52		
а -	-0.44	0.24	-0.42	-0.011	-0.43	0.2	-0.31	0.31	-0.43	-0.1	1	-0.35	0.32		-0.4
LSTAT	0.59	-0.45	0.61	-0.053	0.6	-0.63	0.61	-0.5	0.55	0.38	-0.35	1	-0.8		-0.6
MEDV	-0.54	0.43	-0.56	0.15	-0.51	0.7	-0.46	0.33	-0.54	-0.52	0.32	-0.8	1		
	CRIM	ΖN	INDUS	CHAS	NOX	RM	AĞE	DİS	TÁX	PTRATIO	В	LSTAT	MEDV		

Actual vs Predicted price



NAME:- Pranav Parasar

SIC:- 21BCED77

Email: pranavparasar99@gmail.com