# SemiAuto Regression Report

## Dataset: BOSTON

Generated on: 2025-06-01 02:07:21

### Project Flow


## Table of Contents

# 1. Data Ingestion

This step involves loading and analyzing the original dataset to understand its structure and characteristics.

## Dataset Overview

**Dataset: boston**

**Train samples: 404, Test samples: 102**

**Target column: MEDV**

## Column Types

**Original Columns:**

CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, MEDV

**Numerical Columns:**

ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, MEDV

**Categorical Columns:**

None

**Skewed Columns:**

ZN, CHAS, NOX, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, MEDV

**Normal Columns:**

INDUS, RM

**Columns with Nulls:**

None

**Columns with Outliers:**

ZN, CHAS, RM, DIS, PTRATIO, B, LSTAT, MEDV

## Highly Correlated Features

**ZN:**

- DIS: 0.6644

**INDUS:**

- NOX: 0.7637
- DIS: -0.7080
- TAX: 0.7208

**NOX:**

- INDUS: 0.7637
- AGE: 0.7315

- DIS: -0.7692
- TAX: 0.6680

**RM:**

- MEDV: 0.6954

**AGE:**

- NOX: 0.7315
- DIS: -0.7479

**DIS:**

- ZN: 0.6644
- INDUS: -0.7080
- NOX: -0.7692
- AGE: -0.7479

**RAD:**

- TAX: 0.9102

**TAX:**

- INDUS: 0.7208
- NOX: 0.6680
- RAD: 0.9102

**LSTAT:**

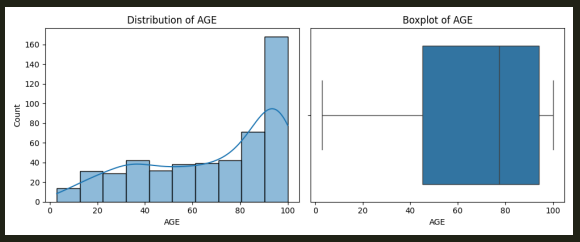
- MEDV: -0.7377

**MEDV:**

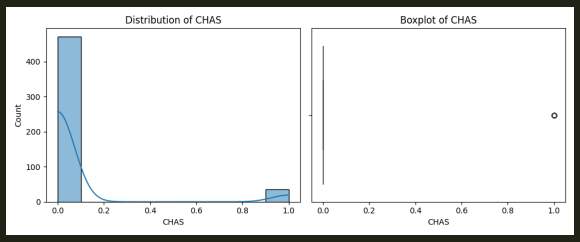
- RM: 0.6954
- LSTAT: -0.7377

# Feature Distributions

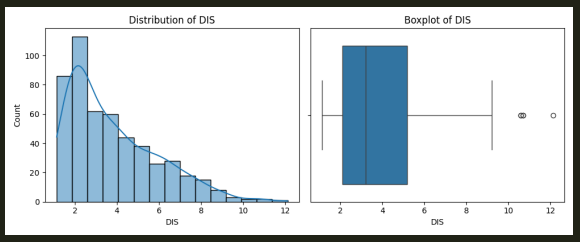
## AGE



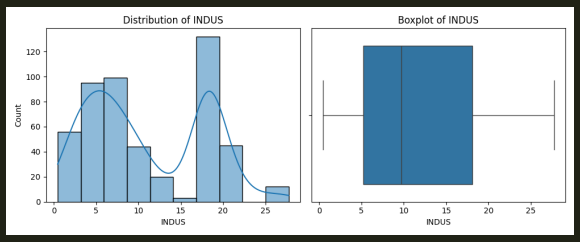

## CHAS



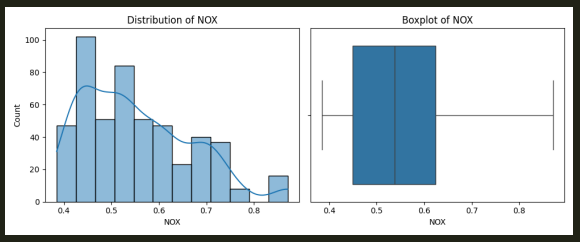

## DIS



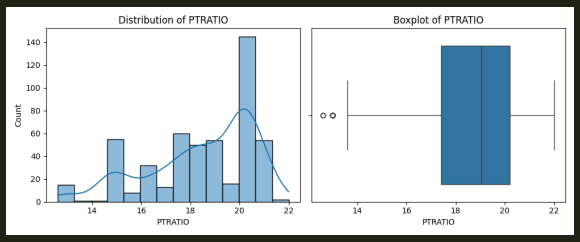

## INDUS



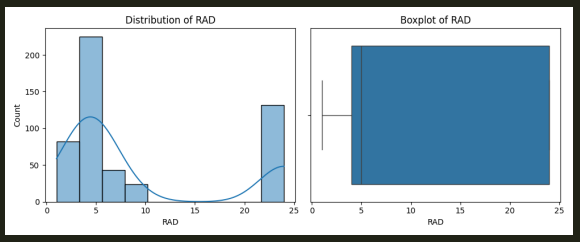

## NOX



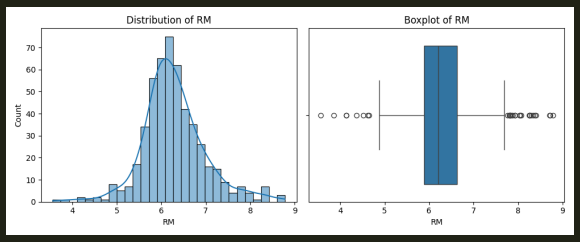

## PTRATIO



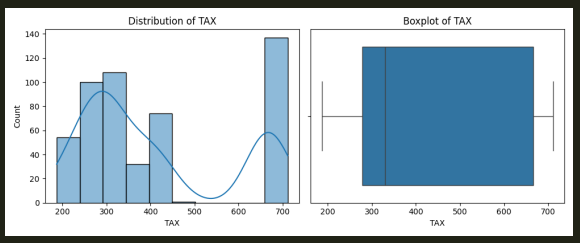

## RAD



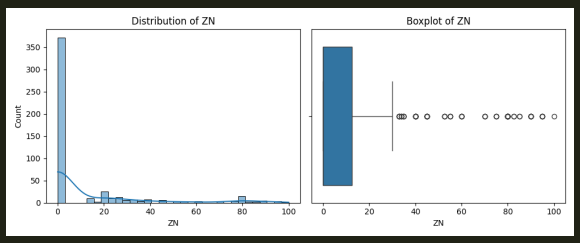

## RM



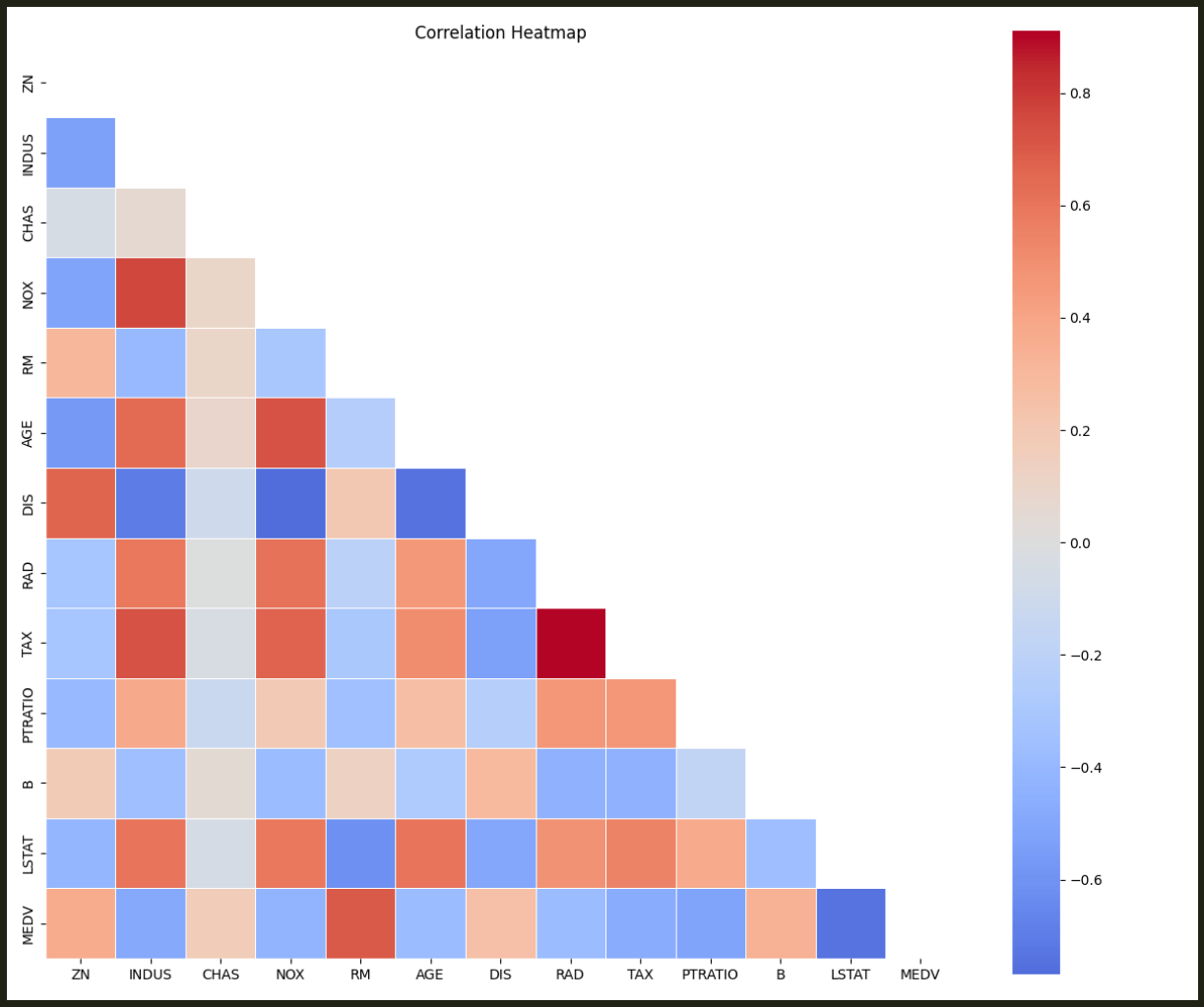

## TAX




## ZN



## Correlation Heatmap



# 2. Data Preprocessing

This step involves cleaning the dataset and preparing it for model training.

## Preprocessed Data Preview

### Training Data Sample (First 5 rows):

| ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO |
|---|---|---|---|---|---|---|---|---|---|
| -0.5267617.. | 1.03323678.. | -1.3554220.. | 0.88174876.. | -1.4280685.. | -2.4129959.. | -2.1745760.. | -0.9191851.. | -2.4259898.. | -7.8341603.. |
| -0.5267617.. | -0.4131595.. | -1.3554220.. | -4.5918952.. | -0.6800865.. | -2.4710876.. | -1.5008121.. | -1.1500485.. | -2.4370104.. | -7.5871578.. |
| -0.4291697.. | -0.7152182.. | -1.3554220.. | -14.247364.. | -0.4020630.. | -2.5077483.. | -1.1869138.. | -1.3571760.. | -2.4371810.. | -8.5884248.. |
| -0.5267617.. | 1.03323678.. | -1.3554220.. | 0.88174876.. | -0.3004503.. | -2.4315892.. | -2.2092622.. | -0.9191851.. | -2.4259898.. | -7.8341603.. |
| -0.5267617.. | -0.4131595.. | -1.3554220.. | -4.5918952.. | -0.8310942.. | -2.4538976.. | -1.6457566.. | -1.1500485.. | -2.4370104.. | -7.5871578.. |

### Test Data Sample (First 5 rows):

| ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO |
|---|---|---|---|---|---|---|---|---|---|
| -0.5267617.. | -1.0071114.. | -1.3554220.. | -6.9734881.. | 0.14128238.. | -2.4330701.. | -1.9656247.. | -1.1165292.. | -2.4376487.. | -8.6389249.. |
| -0.4286824.. | -0.6643910.. | -1.3554220.. | -13.226896.. | 0.62394248.. | -2.4984069.. | -1.5842204.. | -1.1500485.. | -2.4404629.. | -8.4599940.. |
| -0.5267617.. | 2.43316256.. | -1.3554220.. | 0.56077528.. | -0.4698048.. | -2.4106449.. | -2.2807843.. | -1.1500485.. | -2.4252493.. | -7.8630814.. |
| -0.5267617.. | -0.0254207.. | -1.3554220.. | -17.212319.. | -0.3540792.. | -2.5197655.. | -1.3671706.. | -1.1500485.. | -2.4371238.. | -8.1049531.. |
| -0.5267617.. | 1.03323678.. | -1.3554220.. | 6.32246663.. | -0.0266607.. | -2.4214963.. | -2.0659541.. | -0.9191851.. | -2.4259898.. | -7.8341603.. |

# 3. Feature Engineering

This step involves creating new features or selecting the most important ones.

## Feature Engineering Configuration

### Applied Techniques:

Automated Feature Engineering: Yes

SHAP-based Feature Selection: Yes

## Transformed Data Preview

### Transformed Training Data Sample (First 5 rows):

| LSTAT - RM | PTRATIO - RM | LSTAT - ZN | AGE - LSTAT | AGE + LSTAT | LSTAT * PT.. | AGE * LSTAT | B - LSTAT | LSTAT + RAD | PTRATIO * RM |
|---|---|---|---|---|---|---|---|---|---|
| -0.1121800.. | -6.4060918.. | -1.0134868.. | -0.8727473.. | -3.9532445.. | 12.0665544.. | 3.71661358.. | -2.3627198.. | -2.4594337.. | 11.1877182.. |
| -1.1335174.. | -6.9070713.. | -1.2868422.. | -0.6574837.. | -4.2846916.. | 13.7600997.. | 4.48157448.. | -2.0756388.. | -2.9636525.. | 5.15992402.. |
| -1.4293086.. | -8.1863617.. | -1.4022018.. | -0.6763767.. | -4.3391200.. | 15.7285978.. | 4.59261934.. | -2.0682557.. | -3.1885477.. | 3.45308822.. |
| -1.2583740.. | -7.5337099.. | -1.0320627.. | -0.8727647.. | -3.9904136.. | 12.2120806.. | 3.79042067.. | -2.3452278.. | -2.4780096.. | 2.35377655.. |
| -0.9151671.. | -6.7560636.. | -1.2194996.. | -0.7076362.. | -4.2001590.. | 13.2491609.. | 4.28514676.. | -2.1445882.. | -2.8963099.. | 6.30564322.. |

### Transformed Test Data Sample (First 5 rows):

| LSTAT - RM | PTRATIO - RM | LSTAT - ZN | AGE - LSTAT | AGE + LSTAT | LSTAT * PT.. | AGE * LSTAT | B - LSTAT | LSTAT + RAD | PTRATIO * RM |
|---|---|---|---|---|---|---|---|---|---|
| -1.9399268.. | -8.7802072.. | -1.2718827.. | -0.6344257.. | -4.2317146.. | 15.5383543.. | 4.37622817.. | -2.0906403.. | -2.9151737.. | -1.2205279.. |
| -2.6172714.. | -9.0839365.. | -1.5646464.. | -0.5050780.. | -4.4917358.. | 16.8635509.. | 4.98014687.. | -1.8954643.. | -3.1433774.. | -5.2785497.. |
| -1.1573249.. | -7.3932766.. | -1.1003679.. | -0.7835152.. | -4.0377747.. | 12.7942534.. | 3.92243209.. | -2.2640033.. | -2.7771782.. | 3.69411350.. |
| -1.5525462.. | -7.7508738.. | -1.3798638.. | -0.6131399.. | -4.4263911.. | 15.4531110.. | 4.80424947.. | -1.9842375.. | -3.0566741.. | 2.86979607.. |
| -1.6123559.. | -7.8074996.. | -1.1122549.. | -0.7824796.. | -4.0605130.. | 12.8403196.. | 3.96887288.. | -2.2537760.. | -2.5582018.. | 0.20886455.. |

# 4. Model Building

This step involves training the regression model on the transformed data.

## Model Selection

### Selected Model:

CatBoost

Training timestamp: 2025-06-01 02:05:15

# 5. Model Evaluation

This step involves evaluating the performance of the trained model.

## Performance Metrics

### Original Model Performance:

Evaluation timestamp: 2025-06-01 02:06:31

| Metric | Value |
|---|---|
| $R^2$ Score | 0.90503 |
| Explained Variance Score | 0.90621 |
| Mean Squared Error | 6.96439 |
| Root Mean Squared Error | 2.63901 |
| Mean Absolute Error | 1.78011 |
| Mean Absolute Percentage Error | 0.09830 |
| Max Error | 14.08738 |

# 6. Model Optimization

This step involves tuning the hyperparameters of the model to improve performance.

Error: Could not decode hyperparameters file.
Optimization timestamp: 2025-06-01 02:06:31

# 7. Final Evaluation Results

This section presents the final performance of the optimized model.

## Optimized Model Performance

| Metric | Value |
|---|---|
| $R^2$ Score | 0.92211 |
| Explained Variance Score | 0.92212 |
| Mean Squared Error | 5.71206 |
| Root Mean Squared Error | 2.38999 |
| Mean Absolute Error | 1.80317 |
| Mean Absolute Percentage Error | 0.10152 |
| Max Error | 9.52735 |

Evaluation timestamp: 2025-06-01 02:06:31

## Performance Comparison

| Metric | Original Model | Optimized Model | Improvement |
|---|---|---|---|
| $R^2$ Score | 0.90503 | 0.92211 | +1.89% |
| RMSE | 2.63901 | 2.38999 | +9.44% |
| MAE | 1.78011 | 1.80317 | -1.30% |

# Conclusion

Summary of the regression model development and performance.

This report summarizes the development of a regression model to predict MEDV using the boston dataset. A CatBoost regression model was trained and optimized using hyperparameter tuning. The optimization process improved the model's $R^2$ score from 0.90503 to 0.92211, representing a 1.89% improvement.

This automatic report was generated to provide insights into the model development process and performance metrics. It includes details about data preprocessing, feature engineering, model selection, and evaluation results.