Boston Housing Price Analysis

Multilinear Regression Analysis on Housing Prices

Nugrah Nurrohman / 28-08-2025

Executive Summary

Boston Housing Price Analysis

1. Combining socioeconomic (lstat, ptratio, black), environmental (nox, dis), structural (rm, tax, indus), and accessibility (rad, zn, chas, crim) variables provides a more balanced and accurate predictor of housing prices.
2. A cross-validation approach (LOOCV) was applied to evaluate model performance.
3. The model achieved a prediction error rate of ~15, an error of ~15 means the model is, on average, off by about $15,000.
4. There is a significant improvement of error rate of simpler single-variable or lower-degree models (which showed errors in the 20–30+ range).

Context Definition

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA.

The following describes the dataset columns:

- **CRIM** - per capita crime rate by town
- **ZN** - proportion of residential land zoned for lots over 25,000 sq.ft.
- **INDUS** - proportion 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** - proportion of owner-occupied units built prior to 1940
- **DIS** - weighted distances to five Boston employment centres
- **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** - % lower status of the population
- **MEDV** - Median value of owner-occupied homes in $1000's

Methodology & Tools Used

Methods:

a. Multilinear Regression
b. Cross-Validation

Tools:

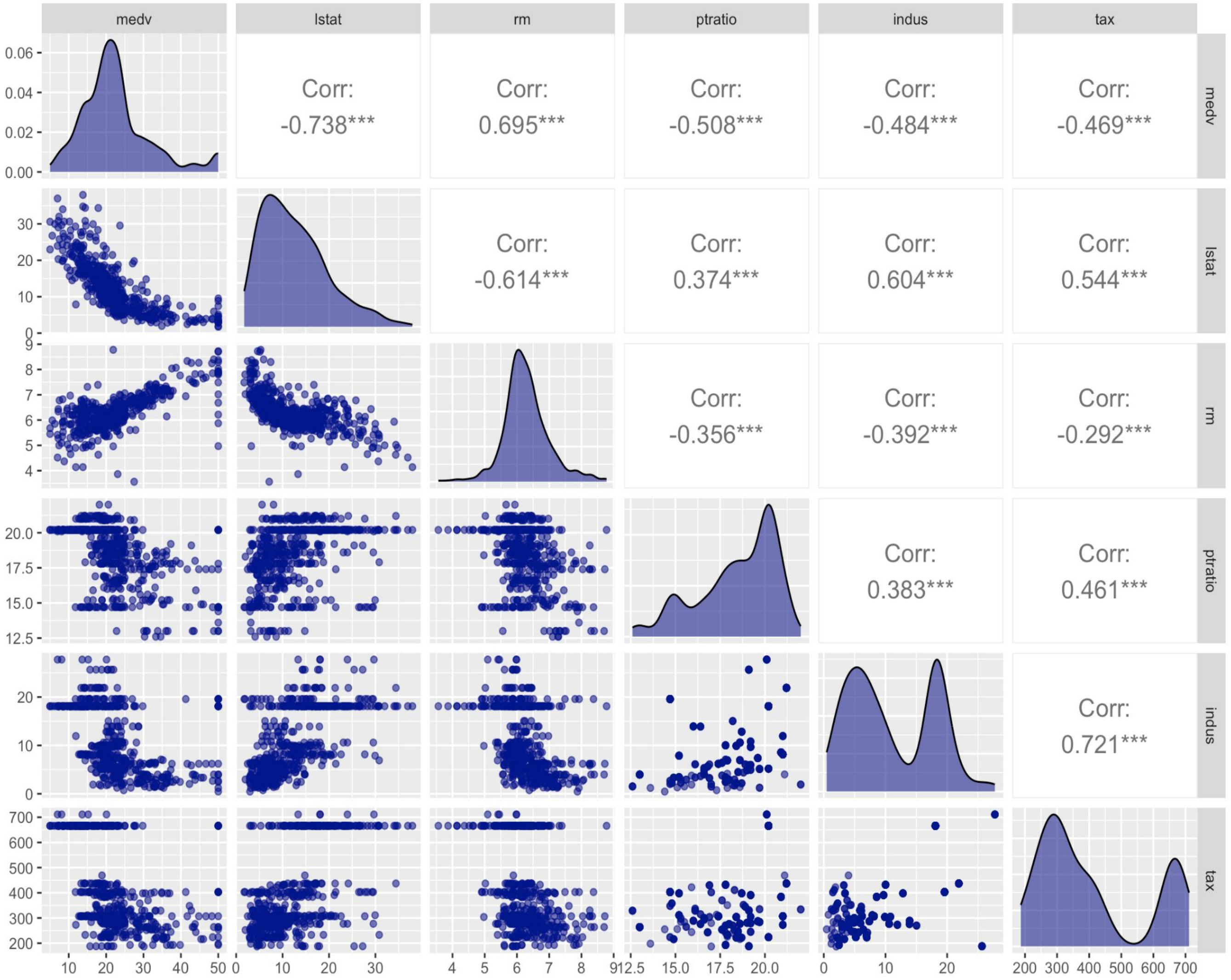
a. R
b. ChatGPT

Key Findings

Most Correlated Variable with House price

Medv(Median Value of Owner-Occupied homes in $1000's)

1. **LSTAT** - % lower status of the population
2. **RM** - average number of rooms per dwelling
3. **PTRATIO** - pupil-teacher ratio by town
4. **INDUS** - proportion of non-retail business acres per town.
5. **TAX** - full-value property-tax rate per $10,000



House Price vs Lower Status Population

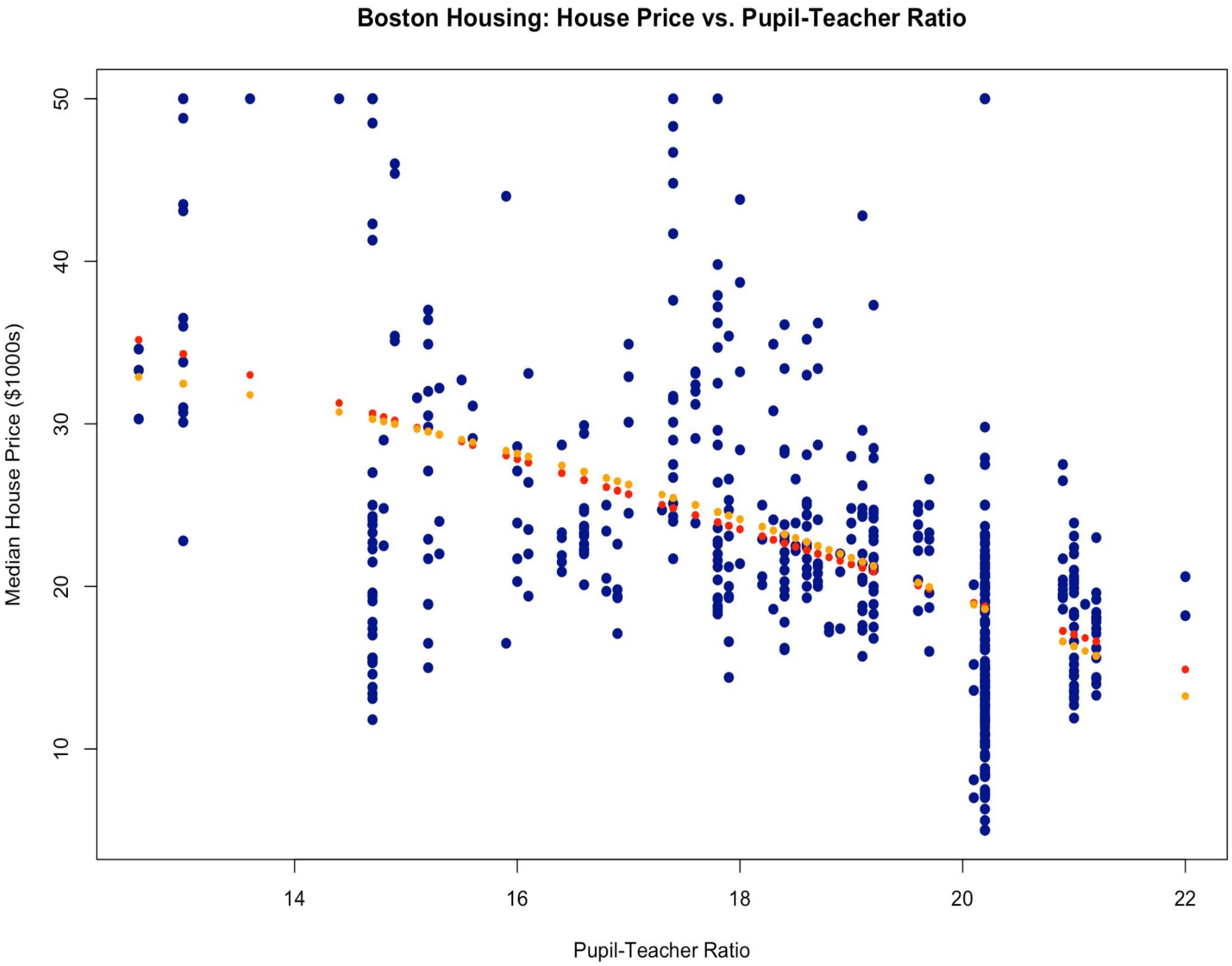
Linear vs Polynomial

Higher lstat level decrease the house price

House Price vs Pupil-Teacher Ratio

Linear vs Polynomial

Higher ptratio (more student than teacher) decrease the house price



House Price vs Number of Rooms

Linear vs Polynomial

More room increase the house price

Interactional model *(Deeper insight)*

-More room in lstat dense area will decrease the price. Individually rm increase the house price but adding interaction, people see bigger house don't fit in the higher lstat dense area.

- High dense lstat area with high ptratio also decrease the price

Variable	Coefficient
lstat:rm	-1.26352
lstat:ptratio	-0.36864

Model Performance

Implementing LOOCV
(Leave- one-out cross validation)
on various model

Single variable Vs Multiple Variable

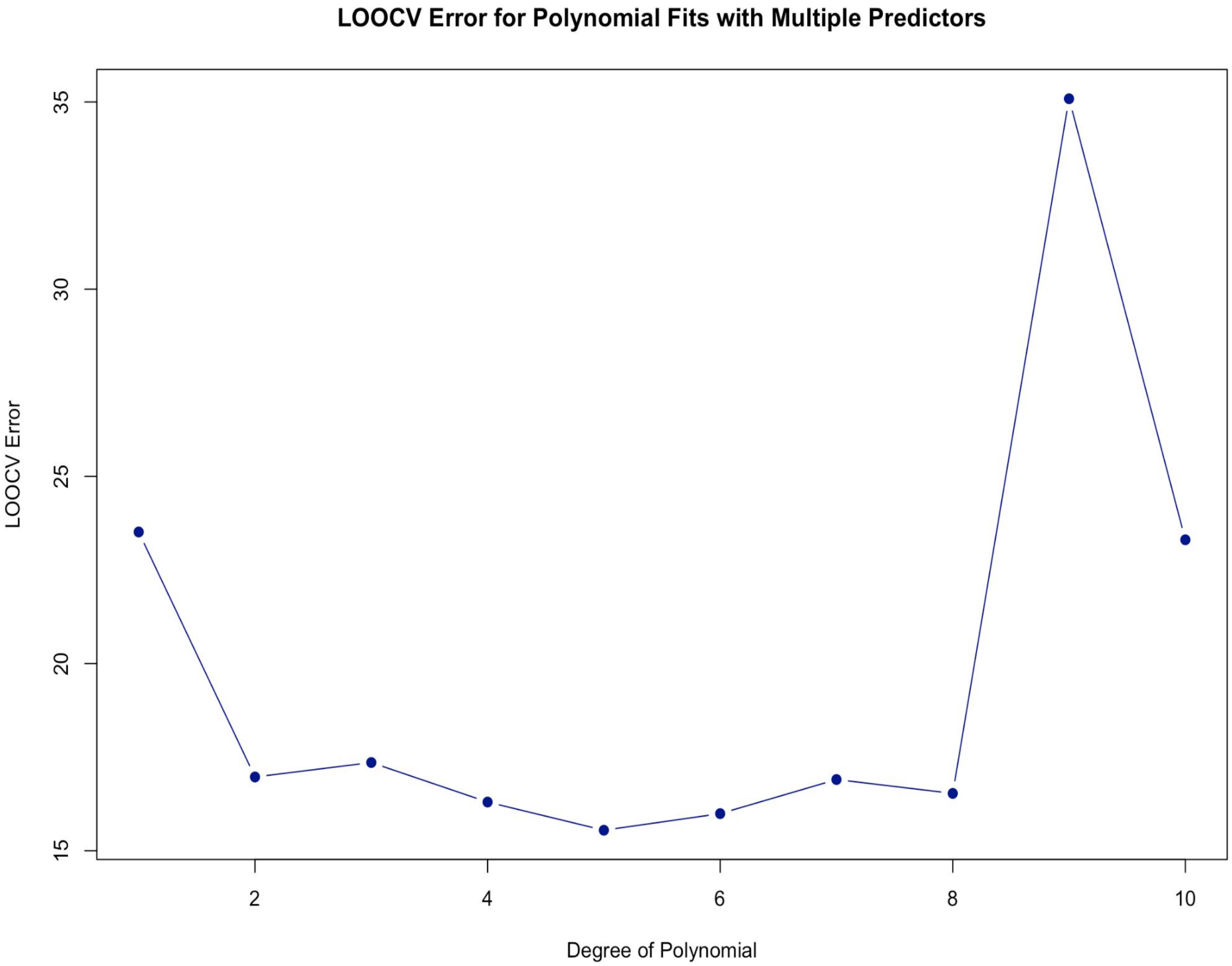
More variables leads to lower prediction error in this dataset.

Variable	Error Rate
lstat	38.89010
ptratio	63.24516
rm	44.21666
lstat+ nox+ptratio+tax+black+rad+dis+rm +nox+chas+zn+crim+indus	23.57196

Multivariate model on various degree

I used polynomial on lstat & rm and left other variable linear.

In the 2 degree model show sharp decline in error rate but the lowest is in the 5 degree and start to overfit in 9 degree



Conclusion

1. Combining socioeconomic (lstat, ptratio, black), environmental (nox, dis), structural (rm, tax, indus), and accessibility (rad, zn, chas, crim) variables provides a more balanced and accurate predictor of housing prices.
2. A cross-validation approach (LOOCV) was applied to evaluate model performance.
3. The model achieved a prediction error rate of ~15, an error of ~15 means the model is, on average, off by about $15,000.
4. There is a significant improvement of error rate of simpler single-variable or lower-degree models (which showed errors in the 20–30+ range).

Appendix

Linear model

```
fit1 <- lm(medv~lstat, data=boston);summary(fit1)

fit4h <- lm(medv~ptratio, data = boston); summary(fit4h)

fit8 <- lm(medv~rm,boston); summary(fit8)

fit11 <- lm(medv~lstat*ptratio*rm,boston); summary(fit11)
```

Loocv with various variable

```
# Fit more varible
glm.fit6 <- glm(medv ~lstat+
nox+ptratio+tax+black+rad+dis+rm+nox+chas+zn+crim+indus
, data = Boston)
coef(glm.fit6)

# LOOCV (default K = n, so it's Leave-One-Out)
cv.error6 <- cv.glm(Boston, glm.fit6)

# Print LOOCV MSE
cv.error6$delta

cv.error6 <- rep(0, 10)

for (d in 1:10) {
  glm.fit6 <- glm(medv ~ poly(lstat, d) + poly(rm, d) + ptratio +
nox+ tax + black + rad + dis + chas + zn + crim ,
              data = Boston)
  cv.error6[d] <- cv.glm(Boston, glm.fit6)$delta[1]
}

cv.error6
```