Boston Housing Dataset Machine Learning Project

By Rohan Sachdeva
PATHS P6 Scholar

Target Variable and Features

- Link: https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset/input
- 506 Observations
- Target: MEDV - Median value of owner-occupied homes in $1000's
- Features:
  - 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

Features (Continued)

- Features:
  - 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

First Five Entries

```
['housing.csv']
```

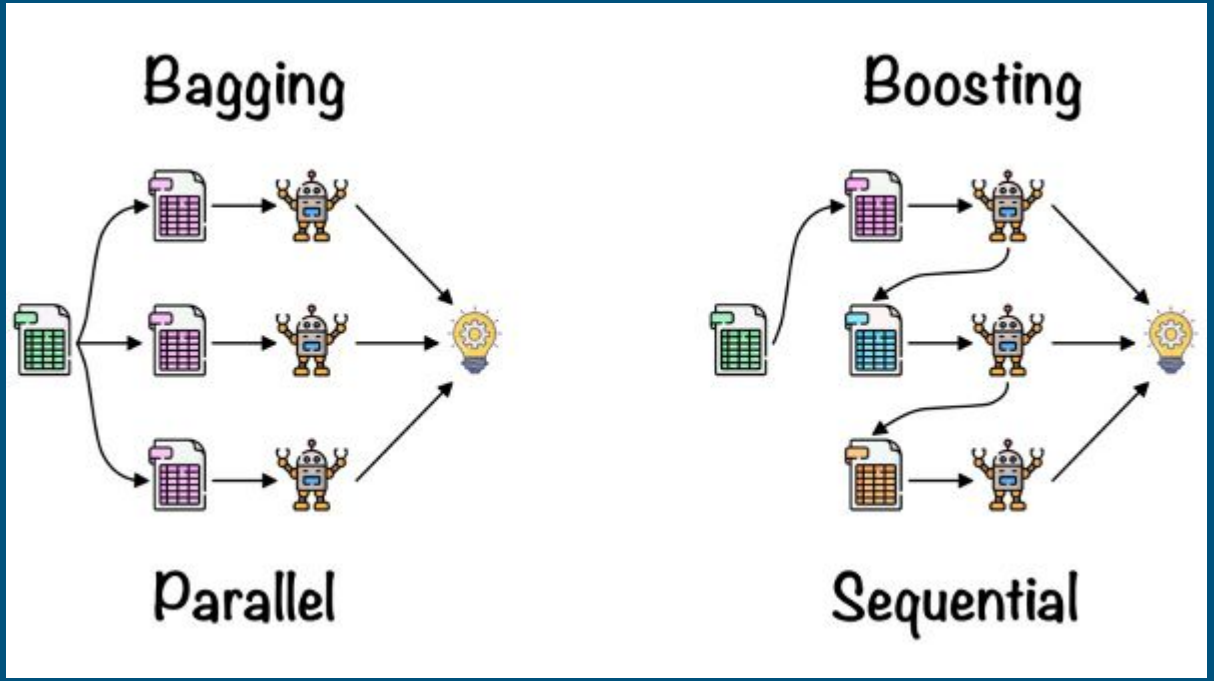
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.90	5.33	36.2

What is Orange?

- Orange Data Mining is a platform which applies a graph structure to utilize machine learning methods and perform Exploratory Data Analysis
  - Provides strong, advanced, unique visualizations, baseline scores for machine learning methods/models against Python code
  - Special tools for Time Series, Socioeconomic, and Geographical Data

Demo: Orange

Bagging and Boosting Algorithms



- Bagging trains multiple independent models to make a prediction and applies some aggregation function (usually mode)
- Chains weak models - each subsequent model covers a subsequent subset of the data with weaker predictions

What is AdaBoost?

- Ensemble technique - combines weak learners to form strong model
  - Assigns higher weights to misclassified examples in each iteration
- Start: Equal weights to all training examples
- Train Weak Learner (slightly better than random guessing)
- Calculate weighted error of the weak learner by summing weights of misclassified examples
- Compute an alpha value - contribution of weak learner to final ensemble so smaller error implies larger alpha
- Increase misclassified weights and decrease properly classified weights
- Normalize updated sample weights to achieve sum of 1
- Repeat above steps for certain iteration number

Evaluating Models - F Test

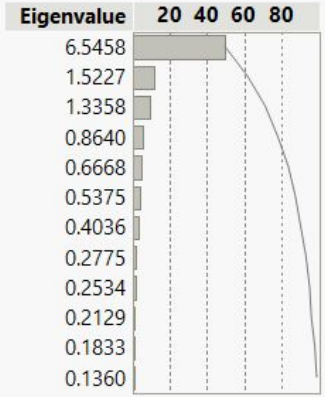
Model	MSE	RMSE	MAE	R2
Gradient Boosting	8.786	2.964	2.078	0.896
AdaBoost	9.427	3.070	2.048	0.888
Random Forest	12.849	3.585	2.339	0.848
Tree	19.472	4.413	3.049	0.769
Neural Network	22.937	4.789	3.448	0.728
Linear Regression	23.370	4.834	3.376	0.723
SGD	24.246	4.924	3.495	0.713
kNN	38.680	6.219	4.352	0.542
SVM	39.153	6.257	3.911	0.536

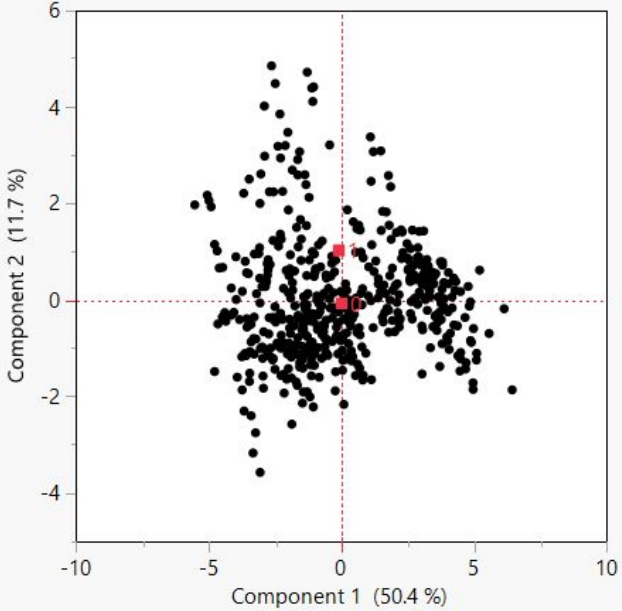
- Based on Orange Test and Score, 8.786 is the MSE for Gradient Boosting and 9.427 for AdaBoost
- We can set up F-Test with H0 = mean performance of Gradient Boosting and AdaBoost are the same; H1 = mean performances are different
- Assuming homogeneity of variance for the performance and simple random sampling
- With this test, I get a P-Value close to 1 and fail to reject the null hypothesis

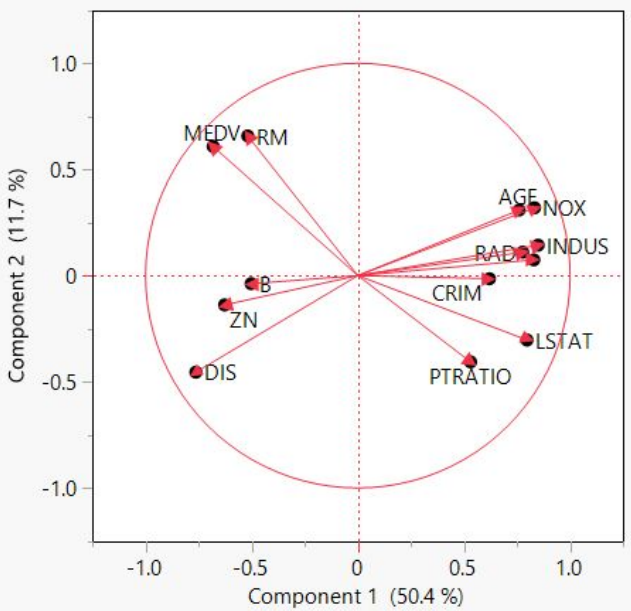
Jupyter Notebook Code

Principal Components: on Correlations

Summary Plots







☑ Label variables