Assignment 03: Tree-based methods

WS24 Data Analytics

Lucas Kook

Goal

In the lab we mostly focused on classification trees and random forests. The goal of this analysis is to get more familiar with regression trees and random forests and to predict housing values in California in the 1990's.

Dataset

The dataset has been obtained from https://www.kaggle.com/datasets/camnugent/california-housing-prices, is available for download on Canvas and can be read into R using `read.csv()`.

The data contains houses from several California districts together with some information about these houses based on the 1990 census data. The following variables are included:

- `longitude` – Coordinates of the house.
- `latitude` – Coordinates of the house.
- `housing_median_age` – Age of the house.
- `total_rooms` – Total number of rooms.
- `total_bedrooms` – Total number of bedrooms.
- `population` – Population of the district.
- `households` – Number of households in the house.
- `median_income` – Median income of inhabitants participating in labor force.
- `median_house_value` – Median value of the house (our response).
- `ocean_proximity` – Distance to the ocean.

Tasks

1. The data contains some missing values. Remove these for the subsequent analysis. Visually compare the marginal distribution of all features and the response before and after removing the missing values.
2. Fit a regression tree to the data (with `median_house_value` as the response and all other features as explanatory variables). Plot the tree and interpret the results.
3. Fit a random forest to the data with permutation importance. Plot and interpret the importance scores and compare them with the results obtained in 2. Do the methods agree or disagree on the important variables? Produce a partial dependency plot for the most important (as measured by permutation importance) feature and interpret the results.
4. Compare the tree, random forest, and linear regression in terms of cross-validated mean squared error in a 10-fold cross validation.

Submission

Note: Only *one* submission per team.

- The assignment should be completed as a notebook containing code and text explanations.
- If you submit an R notebook, please submit the `.html` output together with the `.Rmd` (or `.qmd`) file.

```
{r setup, include=FALSE} set.seed(24101968) knitr::opts_chunk$set(echo =
TRUE, message = FALSE, warning = FALSE,                         error = FALSE,
cache = TRUE)
```

Goal

In the lab we mostly focused on classification trees and random forests. The goal of this analysis is to get more familiar with regression trees and random forests and to predict housing values in California in the 1990's.

Dataset

The dataset has been obtained from https://www.kaggle.com/datasets/camnugent/california-housing-prices and are available for download on Canvas.

The data contains houses from several California districts together with some information about these houses based on the 1990 census data. The following variables are included:

- `longitude` – Coordinates of the house.
- `latitude` – Coordinates of the house.
- `housing_median_age` – Age of the house.
- `total_rooms` – Total number of rooms.
- `total_bedrooms` – Total number of bedrooms.
- `population` – Population of the district.
- `households` – Number of households in the house.
- `median_income` – Median income of inhabitants participating in labor force.
- `median_house_value` – Median value of the house (our response).
- `ocean_proximity` – Distance to the ocean.

Tasks

1. The data contains some missing values. Remove these for the subsequent analysis. Visually compare the marginal distribution of all features and the response before and after removing the missing values.

2. Fit a regression tree to the data (with `median_house_value` as the response and all other features as explanatory variables). Plot the tree and interpret the results.

3. Fit a random forest to the data with permutation importance. Plot and interpret the importance scores and compare them with the results obtained in 2. Do the methods agree or disagree on the important variables? Produce a partial dependency plot for the most important (as measured by permutation importance) feature and interpret the results.

4. Compare the tree, random forest, and linear regression in terms of cross-validated mean squared error in a 10-fold cross validation.

Submission

Note: Only *one* submission per team.

- The assignment should be completed as a notebook containing code and text explanations.
- If you submit an R notebook, please submit the `.html` output together with the `.Rmd` (or `.qmd`) file.