# Decission Trees - House Price Prediction

In this project I consider the California Housing Data (1990) set(https://www.kaggle.com/datasets/harrywang/housing). This data set is created for prediction of median_house_value of California Housings. The data set consists of 10 predictor variables with a sample size of 20433. I take median_house_value as the quantitative response variable. Among the predictors, ocean_proximity is a qualitative variable and treat others as quantitative variables. I consider all the data as training data.

Additionally for all the models, I use 5-Fold cross-validation to compute the estimated test MSE.

## Fit a decision tree to the data and summarize the results.

```
##
## Regression tree:
## tree(formula = median_house_value ~ ., data = housePrice.data)
## Variables actually used in tree construction:
## [1] "median_income"   "ocean_proximity" "longitude"
## Number of terminal nodes:  8
## Residual mean deviance:  5.657e+09 = 1.155e+14 / 20420
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -357300  -45240  -12860       0   32640  408900
```

The Variables actually used in tree construction are “median_income” “ocean_proximity”, and “longitude”. There are 8 nodes and residual mean deviance is 5.657e+09. The distribution of residuals is given below.

```r
summary(sumry$residuals)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -357342  -45242  -12864       0   32636  408944
```

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| -357342 | -45242 | -12864 | 0 | 32636 | 408944 |

Table 1: The distribution of residuals

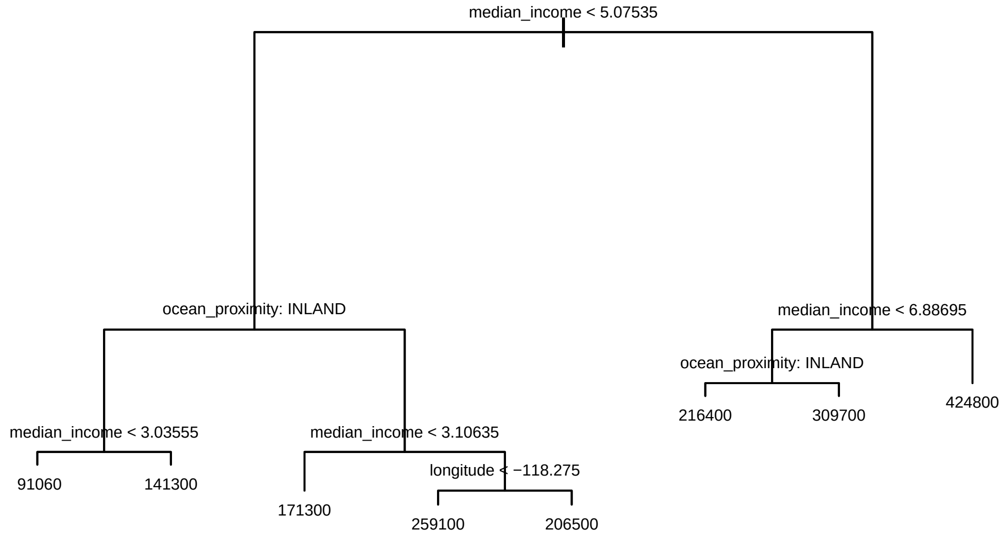


Figure 1: Regression tree for housePrice data

Let $R_j$ be the partitions of the predictor space.

$$R_1 = \{X \mid median\_income < 3.03555, ocean\_proximity = INLAND\}$$
$$R_2 = \{X \mid 3.03555 \leq median\_income < 5.07535, ocean\_proximity = INLAND\}$$
$$R_3 = \{X \mid median\_income < 3.10635, ocean\_proximity = NOT\_INLAND\}$$
$$R_4 = \{X \mid 3.10635 \leq median\_income < 5.07535, ocean\_proximity = NOT\_INLAND, longitude < -118.275\}$$
$$R_5 = \{X \mid 3.10635 \leq median\_income < 5.07535, ocean\_proximity = NOT\_INLAND, -118.275 \leq longitude\}$$
$$R_6 = \{X \mid 5.07535 \leq median\_income < 6.88695, ocean\_proximity = INLAND\}$$
$$R_7 = \{X \mid 5.07535 \leq median\_income < 6.88695, ocean\_proximity = NOT\_INLAND\}$$
$$R_8 = \{X \mid 6.88695 \leq median\_income\}$$

```
library(rpart)
library(caret)
set.seed(1)
K_fold_a<-function(data,k=5){
# Create the folds
folds <- createFolds(data$median_house_value, k = k, list = TRUE, returnTrain = FALSE)

# Initialize a vector to store the evaluation metric values
evaluation_metrics <- c()

# Loop over the folds
for (fold in folds) {
  # Split the data into training and test sets
  train_data <- data[-fold, ]
  test_data <- data[fold, ]

  # Fit the regression tree model on the training data
  fit <- tree(median_house_value ~ ., data = train_data)

  # Predict the target variable on the test data
```

```r
  predictions <- predict(fit, newdata = test_data)

  # Calculate the evaluation metric(s) of interest
  evaluation_metric <- mean((predictions - test_data$median_house_value)^2)  # MSE as an example
  evaluation_metrics <- c(evaluation_metrics, evaluation_metric)
}

# Compute the average evaluation metric across all folds
average_metric <- mean(evaluation_metrics)
return(average_metric)
}

test.MSE<-K_fold_a(data=housePrice.data)
test.MSE
```

```
## [1] 5661830485
```

The test MSE using 5-Fold cross validation is 5661830485.

### Use 5-Fold cross validation to determine whether pruning is helpful and determine the optimal size for the pruned tree.


Figure 2: Plot the estimated test error rate

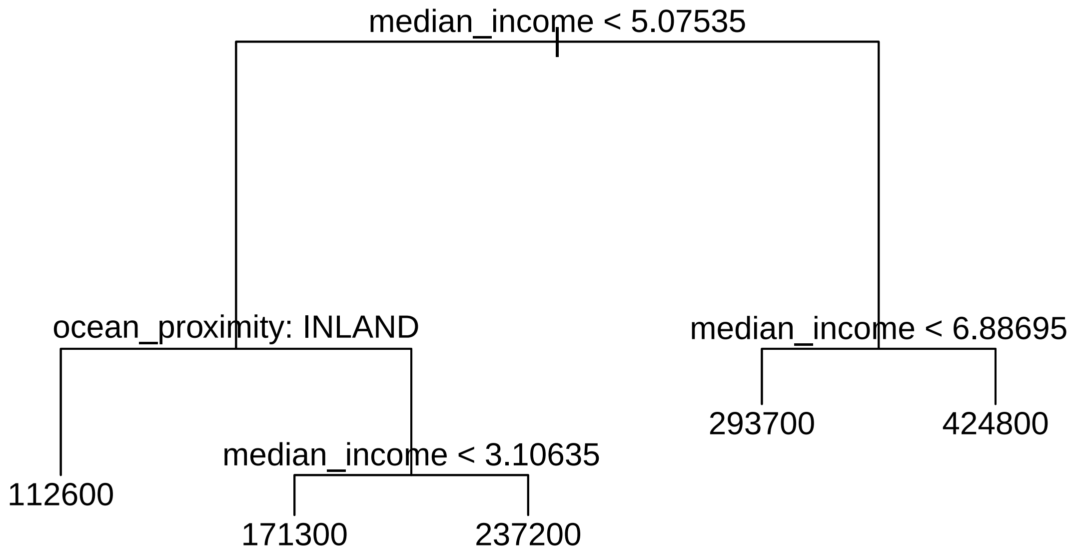


Figure 3: Regression prune Tree for cancer data

```
## [1] 6205063791
```

The pruned tree has five(5) terminal nodes(Figure 2) and the actual used variable in tree construction are "median_income", "ocean_proximity"(See Figure 3) and they are seem to be most important predictors. Using 5-fold cross validation method the test MSE for pruned tree with five terminal nodes is 6205063791. Test MSE is greater than the un-pruned tree .

### Use a bagging approach to analyze the data with $B = 1000$.

```
##                       %IncMSE IncNodePurity
## longitude           86.98483  3.352887e+13
## latitude            78.22694  3.149993e+13
## housing_median_age 118.73843  1.264821e+13
## total_rooms         46.62881  1.186458e+13
## total_bedrooms      67.33324  8.846389e+12
## population          81.19784  1.362355e+13
## households          62.43046  8.240597e+12
## median_income      212.13485  1.050356e+14
## ocean_proximity    125.74147  4.202972e+13
```

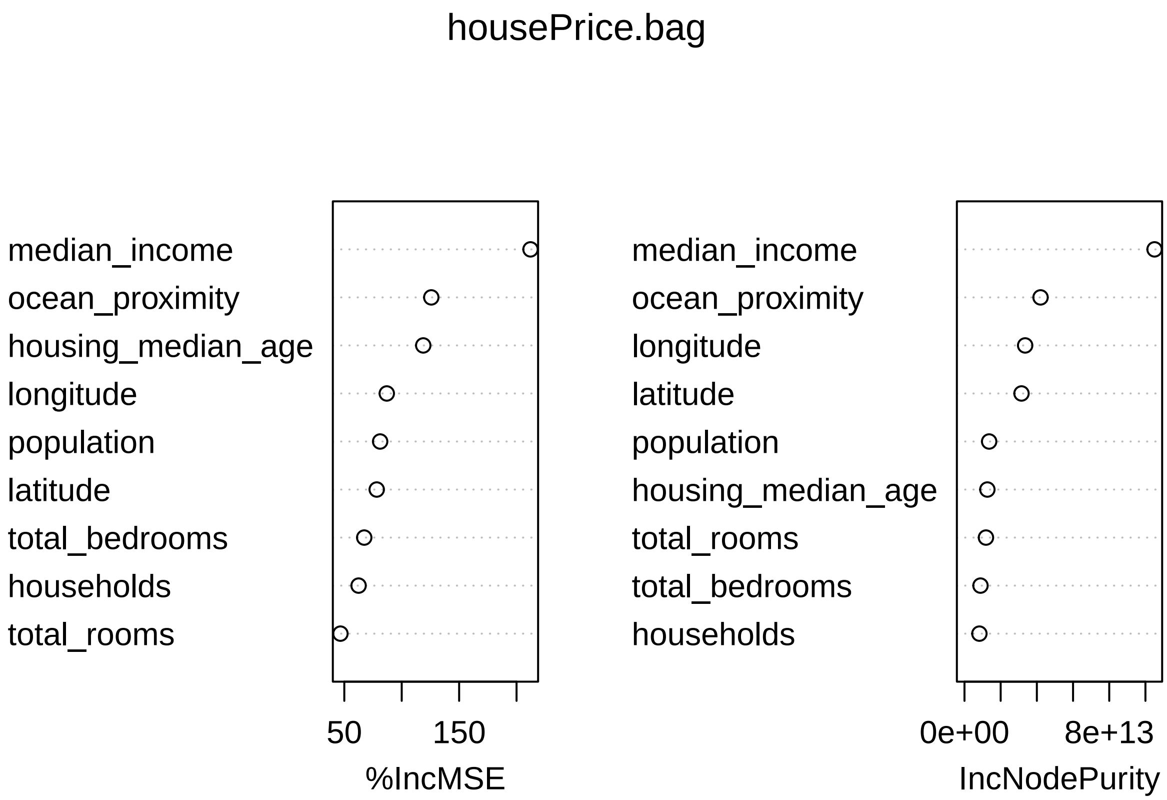


Figure 4: Variable importance measure for each predictor (Bagging)

Using bagging approach with $B = 1000$, the Node purity plot (Figure 4) shows that the variables "median_income"(IncNodePurity=1.050356e+14) and ocean_proximity(IncNodePurity= 4.202972e+13) are the most important predictors. And the test MSE using 5-Fold cross validation method is 2362805450.

### Use a random forest approach to analyze the data with $B = 1000$ and $m \approx p/3$.

```
##                       %IncMSE IncNodePurity
## longitude            86.98483  3.352887e+13
## latitude             78.22694  3.149993e+13
## housing_median_age 118.73843  1.264821e+13
## total_rooms          46.62881  1.186458e+13
## total_bedrooms       67.33324  8.846389e+12
## population           81.19784  1.362355e+13
## households           62.43046  8.240597e+12
## median_income       212.13485  1.050356e+14
## ocean_proximity     125.74147  4.202972e+13
```

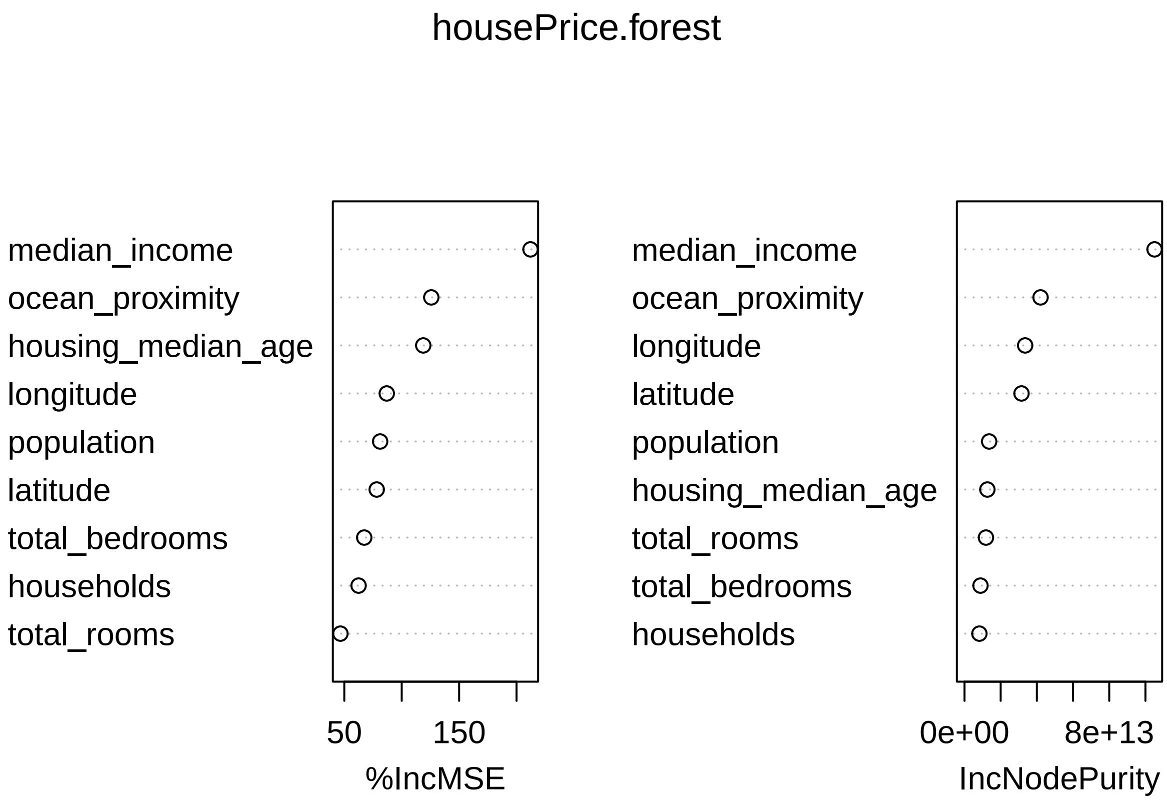


Figure 5: Variable importance measure for each predictor (Random forest)

```
## [1] 2362805450
```

Using random forest approach with $B = 1000$ the Node purity plot (Figure 5) shows that the variables "median_income"(IncNodePurity=1.050356e+14) is most important predictor. And the test MSE using 5-Fold cross validation method is 2362805450.

### Use a boosting approach to analyze the data with $B = 1000$, $d = 1$, and $\lambda = 0.01$.

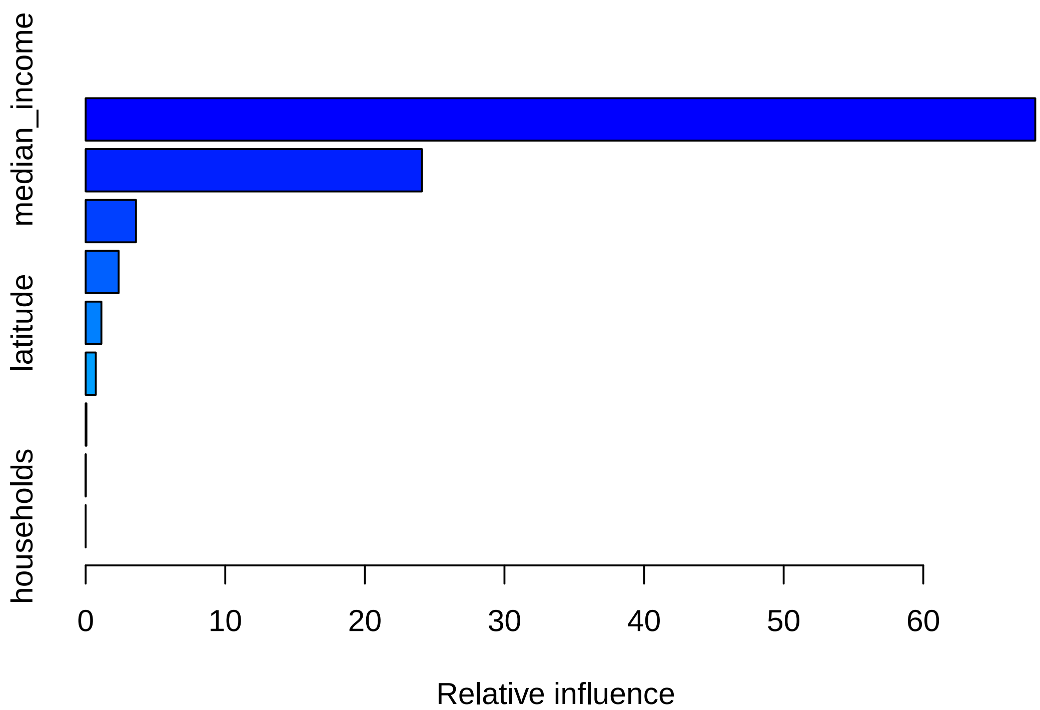


Figure 6: Relative influence Plot

```
##                                   var      rel.inf
## median_income           median_income 68.024004517
## ocean_proximity       ocean_proximity 24.089808146
## longitude                   longitude  3.604430754
## housing_median_age housing_median_age  2.362567762
## latitude                     latitude  1.127838857
## total_bedrooms         total_bedrooms  0.727908889
## total_rooms               total_rooms  0.054227604
## population                 population  0.009213471
## households                 households  0.000000000
```

```
## [1] 4886484198
```

Using boosting approach with $B = 1000$, $d = 1$ and $\lambda = 0.01$, according to the Relative influence plot (Figure 6) it shows that the variables " median_income" (rel.inf=68.024004517) and "ocean_proximity" (rel.inf= 24.089808146) are most important predictors. And the test MSE using 5-Fold cross validation method is 4886484198.

### Finnally I compare the results from the various methods.

| | un-pruned tree | pruned tree | bagging | random-forest | boosting |
|---|---|---|---|---|---|
| Test MSE | 5661830485 | 6205063791 | 2362805450 | 2362805450 | 4886484198 |

Table 2: Test MSE for different approches

When consider the four different approaches discussed above, pruned tree approach gives large test MSE(6205063791) and bagging approach gives the small test MSE(2307717850). So bagging approach should be recommended to analyse California Housing Data.