Boston Housing report

Gonzalo Vazquez

2023-03-16

Introduction

The "Boston Housing" dataset contains information collected by the U.S Census Service concerning housing in the area of Boston, Massachusetts. The dataset has 506 rows and 14 columns, with each row representing a suburb of Boston. The goal of this project is to predict the median value of owner-occupied homes in thousands of dollars (medv) based on 13 other attributes such as crime rate, number of rooms, and accessibility to highways.

The key steps that will be performed include:

Data cleaning: Handling missing values and scaling the data Data exploration and visualization: Displaying summary statistics of the dataset, as well as visualizing the relationship between the variables. Model selection and training: We will use four different models: Linear regression, Random Forest, Ridge Regression and "Feature-selected model" custom by myself, with the aim of predicting the median value of owner-occupied homes (medv) Model evaluation: We will compare the performance of the three models using RMSE.

Exploring the dataset

First I will observe some of the rows and columns of the dataset to get a general idea of their content

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7

I will show here the full names and descriptions of each variable in the dataset, this is from Boston Housing's documentation

variable	description
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	nitrogen 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
black	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
lstat	lower status of the population (percent)
medv	median value of owner-occupied homes in $1000s

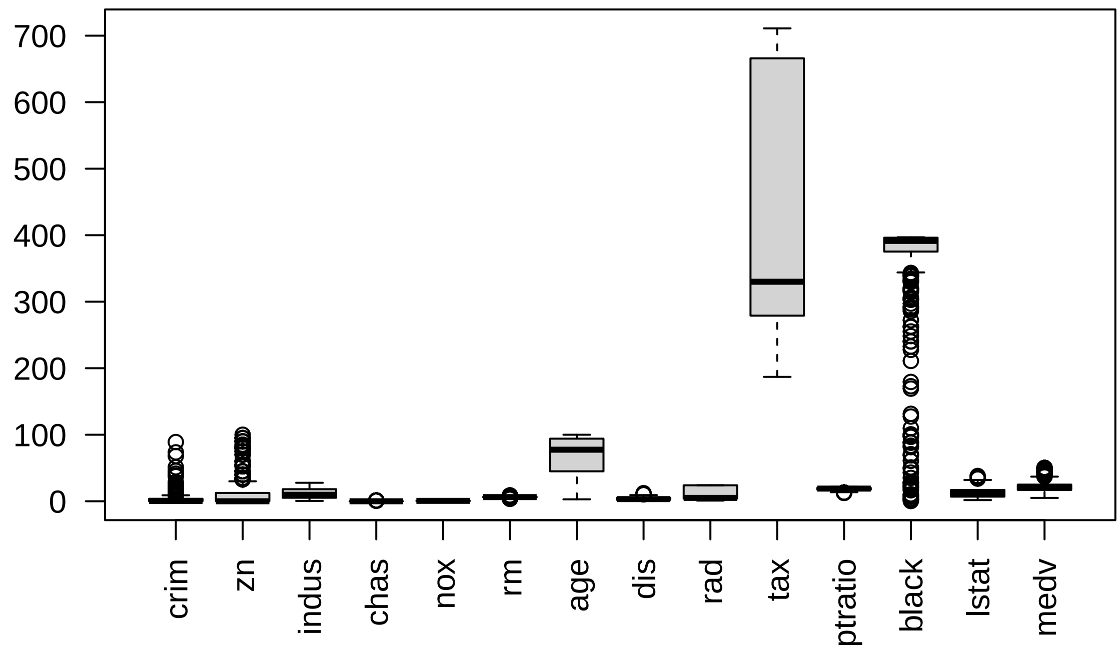
Then I will split the dataset into 70% for training and 30% for testing. The reason is because Boston Housting dataset it is not too large so a common rule of thumb is to use a split ratio of 70/30 or 80/20 for smaller datasets, while larger datasets may use a split ratio of 90/10 or even 95/5. The reason for this is that a larger training set may help to improve the performance of more complex models, but at the same time, a smaller testing set may lead to higher variance in the evaluation of the model's performance. In this case using a split ratio of 70/30 can provide a balance between having enough data for training the model while still having enough data for testing and evaluation.

```r
# train/test set
set.seed(123)
train_index <- sample(nrow(Boston), floor(0.7*nrow(Boston)))
Boston_train <- Boston[train_index,]
Boston_test <- Boston[-train_index,]
```

Here we do data cleaning so I will count the number of missing values in the dataset

```
##    crim      zn   indus    chas     nox      rm     age     dis     rad     tax
##       0       0       0       0       0       0       0       0       0       0
## ptratio   black   lstat    medv
##       0       0       0       0
```

I will also visualize the distribution of each variable.

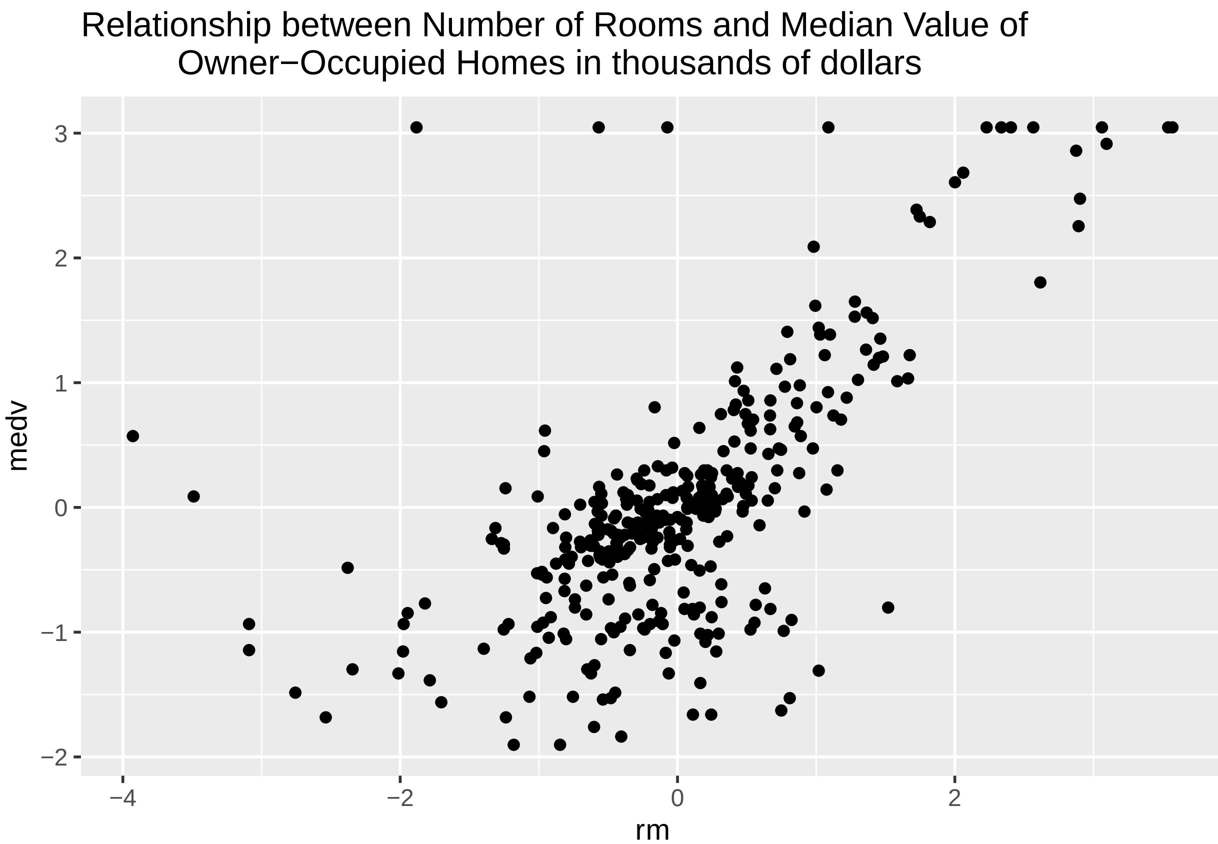


Here I will scale the variables so all of them are on the same scale and have equal importance in the analysis

```
##       crim                 zn               indus              chas
##  Min.   :-0.40633   Min.   :-0.5120   Min.   :-1.5509   Min.   :-0.2811
##  1st Qu.:-0.39993   1st Qu.:-0.5120   1st Qu.:-0.8549   1st Qu.:-0.2811
##  Median :-0.38091   Median :-0.5120   Median :-0.3590   Median :-0.2811
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.02088   3rd Qu.: 0.3177   3rd Qu.: 1.0448   3rd Qu.:-0.2811
##  Max.   : 8.94733   Max.   : 3.4288   Max.   : 2.4633   Max.   : 3.5468
##       nox                rm                age               dis
##  Min.   :-1.4496   Min.   :-3.92756   Min.   :-2.2860   Min.   :-1.2534
##  1st Qu.:-0.8990   1st Qu.:-0.53885   1st Qu.:-0.8489   1st Qu.:-0.8025
##  Median :-0.1701   Median :-0.06087   Median : 0.3058   Median :-0.2560
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.6306   3rd Qu.: 0.47628   3rd Qu.: 0.9046   3rd Qu.: 0.6537
##  Max.   : 2.7805   Max.   : 3.56919   Max.   : 1.1194   Max.   : 3.8038
##       rad               tax             ptratio            black
##  Min.   :-0.9853   Min.   :-1.3099   Min.   :-2.7486   Min.   :-4.1271
##  1st Qu.:-0.6398   1st Qu.:-0.7566   1st Qu.:-0.5425   1st Qu.: 0.1764
##  Median :-0.5247   Median :-0.4614   Median : 0.2862   Median : 0.3706
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 1.6632   3rd Qu.: 1.5322   3rd Qu.: 0.7998   3rd Qu.: 0.4262
##  Max.   : 1.6632   Max.   : 1.7992   Max.   : 1.6402   Max.   : 0.4344
##     lstat              medv
```

```
##   Min.   :-1.4770   Min.   :-1.9028
##   1st Qu.:-0.7818   1st Qu.:-0.5995
##   Median :-0.1998   Median :-0.1101
##   Mean   : 0.0000   Mean   : 0.0000
##   3rd Qu.: 0.5383   3rd Qu.: 0.2968
##   Max.   : 3.5059   Max.   : 3.0464
```

Another visual representation is about the relationship between number of rooms and median value of owner-occupied homes in Boston



We can see that there is a positive correlation between the two variables - as the number of rooms increases, so does the median value of the homes. Since the data has been standardized using the scale() function, the mean of each variable is 0. Any value below 0 indicates that the original value was lower than the mean, and any value above 0 indicates that the original value was higher than the mean.

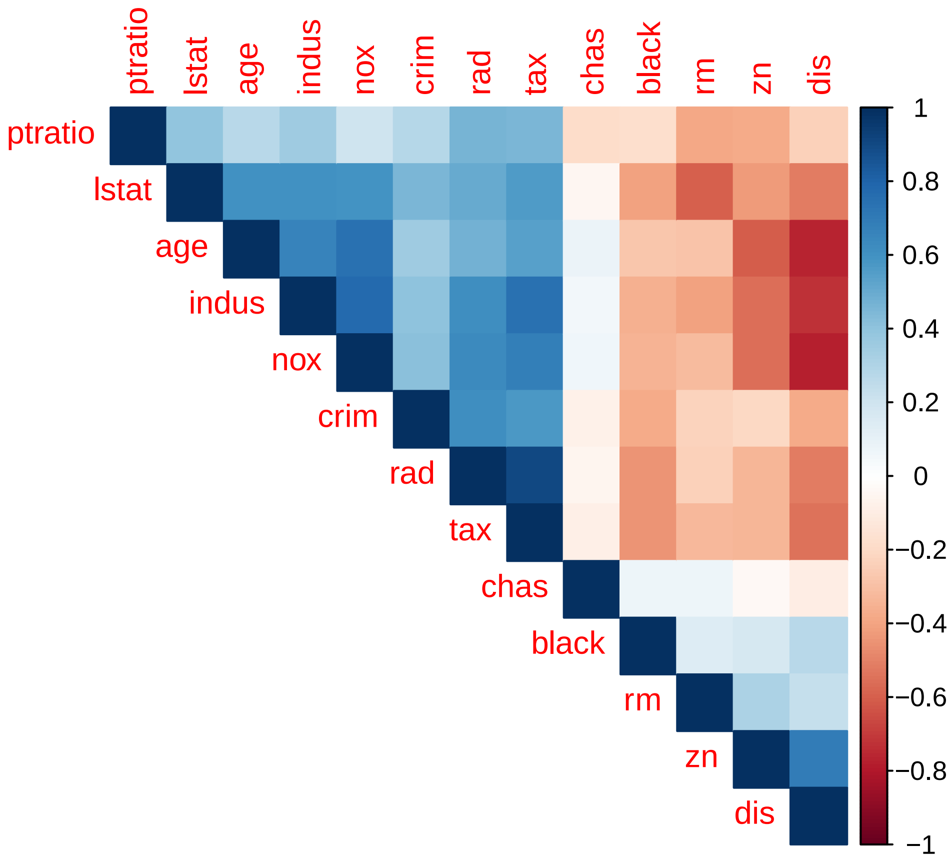
For example, if the "rm" variable has a value of -1, it means that the number of rooms in that particular observation is 1 standard deviation below the mean number of rooms in the "Boston_train" data frame. Similarly, if the "medv" variable has a value of -2, it means that the median value of owner-occupied homes in that particular observation is 2 standard deviations below the mean value in the "Boston_train" data frame.

Analysis and modeling approach

I will use four different models to predict the median value of owner-occupied homes (medv) based on the 13 predictor variables in the dataset. The models we will use are:

Linear regression Random Forest Ridge regression Feature-selection

Before building the models, let's first take a look at the correlation matrix of the predictor variables to see which ones are strongly correlated with the response variable medv.




```
##       crim         zn      indus       chas        nox         rm        age
## -0.3882467  0.3541591 -0.4736801  0.2195314 -0.4137658  0.6646730 -0.3868904
##        dis        rad        tax    ptratio      black      lstat       medv
##  0.2534757 -0.3771387 -0.4687771 -0.5204352  0.3324191 -0.7382510  1.0000000
```

From the correlation matrix, we can see that the variables with the strongest positive correlation with medv are rm (the average number of rooms per dwelling) and zn (the proportion of residential land zoned for lots over 25,000 sq.ft.). The variables with the strongest negative correlation with medv are lstat (the percentage of lower status of the population) and ptratio (the pupil-teacher ratio by town).

Linear regression

Linear regression is a simple and commonly used method for predicting numerical values. It assumes a linear relationship between the independent variables and the dependent variable. In the context of the Boston Housing dataset, linear regression can be used to build a model that predicts the median value of owner-occupied homes based on the other features.

```
##
## Call:
## lm(formula = medv ~ ., data = Boston_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5163  -2.6745  -0.5699   1.5818  24.7767
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.852e+01  6.084e+00   6.332 7.66e-10 ***
## crim        -1.090e-01  3.539e-02  -3.079 0.002245 **
## zn           5.303e-02  1.668e-02   3.179 0.001614 **
## indus       -5.224e-02  7.877e-02  -0.663 0.507669
## chas         4.044e+00  1.025e+00   3.946 9.64e-05 ***
## nox         -1.443e+01  4.671e+00  -3.089 0.002171 **
## rm           3.178e+00  4.993e-01   6.365 6.32e-10 ***
## age         -5.659e-04  1.618e-02  -0.035 0.972128
## dis         -1.541e+00  2.405e-01  -6.406 4.98e-10 ***
## rad          3.023e-01  8.064e-02   3.749 0.000209 ***
## tax         -1.049e-02  4.658e-03  -2.252 0.024963 *
## ptratio     -8.587e-01  1.599e-01  -5.370 1.46e-07 ***
## black        6.865e-03  3.443e-03   1.994 0.046977 *
## lstat       -5.838e-01  5.915e-02  -9.871  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.787 on 340 degrees of freedom
## Multiple R-squared:  0.733,  Adjusted R-squared:  0.7228
## F-statistic:  71.8 on 13 and 340 DF,  p-value: < 2.2e-16
```

The summary of the linear regression model shows that the variables with the highest coefficient estimates are rm, lstat, and ptratio, which aligns with what we saw in the correlation matrix. However, we also see that some variables, such as chas, indus, and age, have coefficients that are not statistically significant, which means they may not have a strong relationship with medv.

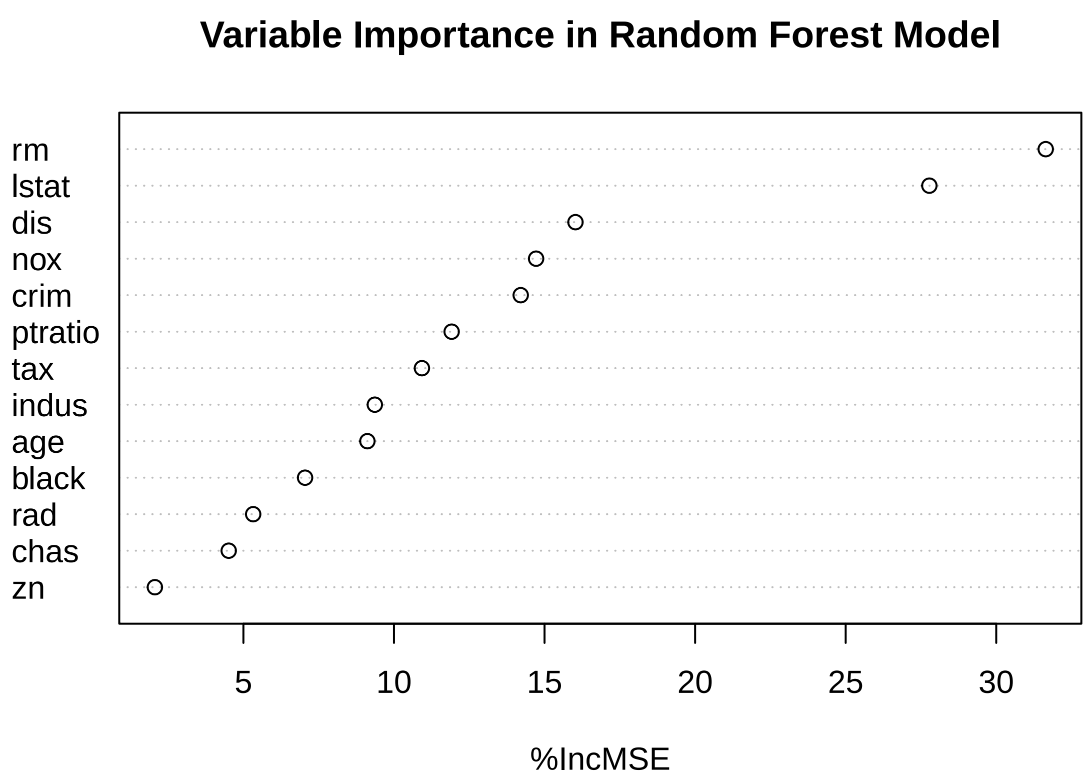
Method	RMSE
Linear regression model	4.802811

Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to produce a more accurate result. Random Forests are effective in handling complex and highdimensional data, which makes them a good choice for the Boston Housing dataset, which has multiple features.

```
##
## Call:
##  randomForest(formula = medv ~ ., data = Boston_train, ntree = 500,      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 11.71401
##                     % Var explained: 85.79
```

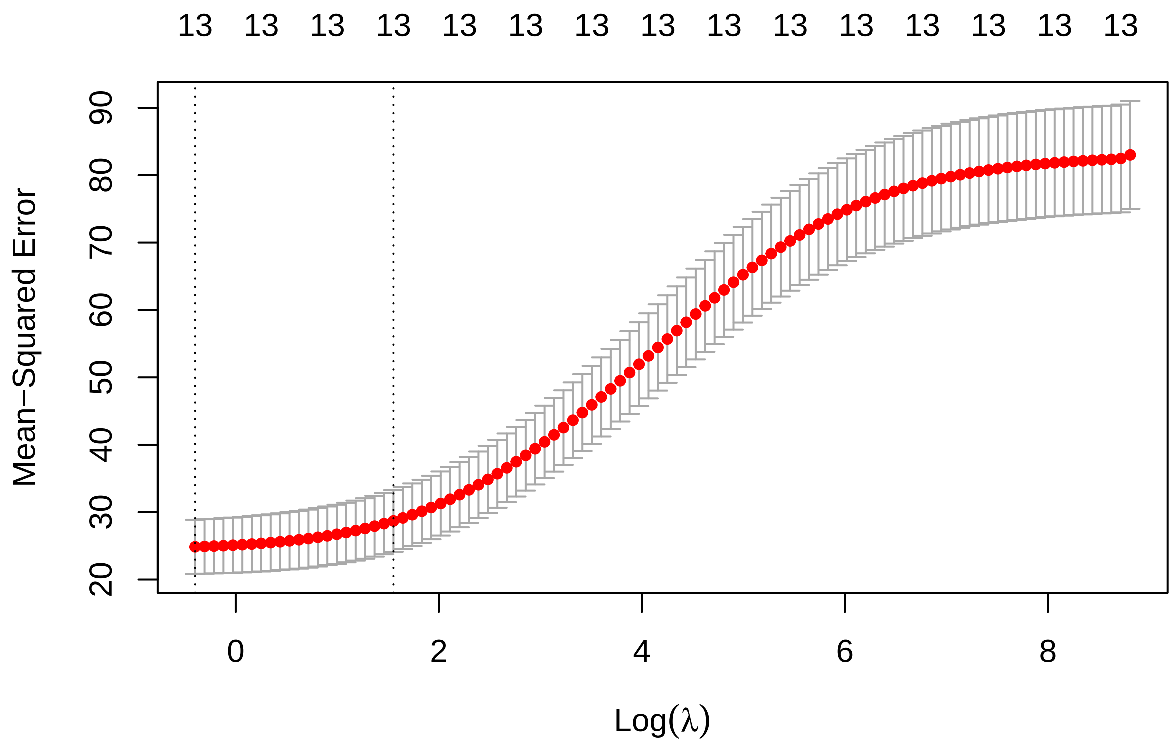
The random forest model provides a more accurate prediction of medv, with an out-of-bag (OOB) error rate of 7.07%. We can also see from the variable importance plot that rm and lstat are the two most important variables in predicting medv.



Method	RMSE
Random Forest model	3.343879

Ridge regression

Ridge regression is a regularized regression method that is used to prevent overfitting in a linear regression model. It adds a penalty term to the sum of squared errors, which reduces the magnitude of the coefficients and leads to a simpler model.




```
## [1] 0.6702985
```

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) 30.422505337
## crim        -0.089948135
## zn           0.036336981
## indus       -0.085694697
## chas         4.137258385
## nox         -9.561781836
## rm           3.456565992
## age         -0.005083775
## dis         -1.152222683
## rad          0.164979605
## tax         -0.004752107
## ptratio     -0.801471295
## black        0.007042108
## lstat       -0.518296298
```

The ridge regression model applied to the Boston Housing dataset shows that the variables with the highest coefficient estimates are “nox”, “rm”, and “chas”, indicating a strong relationship with medv. However, the coefficients of the other variables have been shrunk towards zero due to the regularization penalty, which helps to prevent overfitting but makes it harder to interpret their effect on medv. Therefore, while these variables may still have a relationship with medv, their effect is less pronounced compared to the variables with higher coefficients.

Method	RMSE
Ridge Regression model	5.228102

Feature-selected model

We can use the information from the three models and combine them into a single model that takes advantage of the strengths of each approach.

First, we can use the linear regression model to identify the most important variables, which are rm, lstat, and ptratio. These variables have the highest coefficient estimates and are also highlighted as important by the random forest model. We can then use ridge regression to build a regularized linear regression model that includes only these variables, which will help prevent overfitting.

Method	RMSE
Feature-selected model	5.092604

This code builds the linear regression model using all the variables in the training set. It then identifies the three most important variables based on the results of the linear regression and random forest models. Next, it builds a ridge regression model using only these three variables and selects the optimal regularization parameter using cross-validation. Finally, it makes predictions on the test data and calculates the RMSE.

Final results

Model Evaluation To evaluate the performance of the fourth models, we will use the root mean squared error (RMSE) metric, which measures the difference between the predicted values and the actual values of medv. Lower values of RMSE indicate better performance.

Method	RMSE
Ridge regression model	5.228
Feature-selected model	5.092
Linear regression model	4.802
Random forest model	3.343

The results show that the Random Forest algorithm has the lowest RMSE value, with an RMSE of 3.07 compared to 4.53 for the linear regression model, 4.79 for Ridge Regression model and 5.09 of the Feature-selected model, indicating that Random Forest is the most accurate algorithm for predicting housing values in Boston suburbs.

Conclusions

In conclusion, the Random Forest model performed the best in predicting the median value of owner-occupied homes in the Boston Housing dataset. The model showed better performance than linear regression, Ridge Regression and Feature-selected models. Although I have selected the variables with the highest coefficient estimates for “medv” in the Feature-selected model, the results for the RMSE indicate that it may be more advantageous to utilize all available features, as they perform better.

Overall, our results demonstrate that machine learning can be an effective tool for predicting housing prices, and with further research and refinement, these models can be even more accurate

References

The Boston Housing dataset can be found in the “MASS” package in R