머신러닝 기반 데이터 분석

다중회귀분석과 랜덤포레스트 기법을 활용한 Boston 주택 가격 예측 및 비교

2023.03.27

B2 팀

서영석, 박용태, 이현호, 전국림

<목차>

1. 서론

1) 데이터 분석 배경

과거부터 현재까지 주택가격은 끊임없이 오르고 내려갔다를 반복하고 있다. 이러한 현상에 대해 다양한 예측기법을 사용하여 여러가지의 변수들이 미국 매사추세츠주의 주도인 보스턴의 주택가격에 어떤 영향을 미치는지 분석 및 예측 후 기법 간에 차이가 있는지 알아볼 것이다.

예측기법에는 인공 신경망 분석, 의사결정트리, 서포트 벡터머신(회귀), PLS(Partial Least Squares), 앙상블기법(랜덤포레스트 등), 선형회귀, 확정된 회귀분석(ex 다항회귀, 비선형 회귀, 벌점화 회귀 등) 등이 있다. 이 중 다중회귀분석 기법과 랜덤포레스트 기법을 사용하였다.

2) 데이터 분석 설명

이 보고서에서는 R 을 활용하여 미국 Boston 주택 가격에 대한 데이터를 분석할 예정이다. 데이터셋은 mlbench 패키지 내 BostonHousing 데이터셋을 사용하였다.

해당 데이터는 1970 년 506 명을 대상으로 한 인구 조사에서, 보스턴 지역에 대한 주택 데이터이다. 데이터셋 내의 변수명의 의미를 아래 표와 같이 정리하였다.

변수	변수 정의
Crim	도시별 1 인당 범죄율
Zn	25,000 평방 피드 이상의 주거 토지 비율
Indus	도시당 비소매 업종 비율
Chas	찰스강에 대한 더미변수(강의 경계 1, 아닐시 0)
Nox	10ppm 당 일산화질소 농도
Rm	주택 1 가구당 평균 방의 개수
Age	1940 년 이전에 건축된 소유주택 비율
Dis	5 개의 보스턴 직업센터까지의 접근성 지수
Rad	방사형 고속도록 접근성 지수
Tax	10,000 달러 당 재산세율
Ptratio	마을별 학생/교사 비율
B	1000(Bk-0.63)^2 (Bk : 자치시별 흑인의 비율)
Lstat	모집단의 하위계층 비율(%)
Medv	보인 소유의 주택 가격(중앙값), 단위 : 1000$

3) 데이터 정제

본론에 들어가기 전, 데이터셋을 가져오는 과정이다. 먼저 [install.packeges("mlbench")]를 실행하여 패키지를 다운 받은 후 [library(mlbench)]를 실행하여 library 를 부착한다.

이후 [data(BostonHousing)]를 실행하여 BostonHousing 데이터셋을 로드한다.

해당 부분에 대한 코드는 아래와 같다. 아래에서 데이터셋의 정보를 'bostonData' 라는 변수에 담았다. 데이터의 기본적인 이해를 돕기 위해 기술통계량 및 데이터 구성을 summary 함수와 dim 함수를 통해 확인하였다.

```
#install.packages("mlbench")
library(mlbench)
library(dplyr)

# 데이터 불러오기
data(BostonHousing)

# BostonHousing 데이터셋 정보 확인
?BostonHousing

# BostonHousing 데이터셋 변수에 담기
bostonData <- BostonHousing

# 기술통계량 확인
summary(BostonHousing)

# 데이터 구성 확인
dim(BostonHousing) # 506 명 대상, 14 개 수치형 변수
```

변수 'bostonData'에는 506 명 대상, 14 개의 수치형 변수로 이루어져있다.

2. 본론

1) 다중회귀분석 기법을 사용한 데이터 분석

각각의 변수들이 Boston 주택가격에 어떠한 영향을 미치는지에 대하여 확인하고자 한다. 따라서 medv 변수를 종속변수로 설정하고 나머지 변수를 독립변수로 설정했다.

첫번째로는 다중회귀분석을 통해 결과값을 예측하고자 한다.

다중회귀분석을 실시 하기 전 다중공선성 문제 확인과 샘플링을 실시하였다.

```
vif(result.lm)
crim : 1.79 / zn : 2.29 / indus : 3.99 / chas : 1.07 / nox : 4.39 / rm : 1.93 / age : 3.1
dis : 3.95 / rad : 7.48 / tax : 9.00 / ptratio : 1.79 / b : 1.34 / lstat : 2.94
```

해당 결과 vif 함수를 사용해 다중공선성이 모두 10 이하로 문제 없는 것을 확인 할 수 있다.

```
set.seed(1234)
idx <- sample(1:nrow(bostonData), nrow(bostonData) * 0.7)
bo_tr <- bostonData[idx, ]
dim(bo_tr) # 506 개 * 0.7 = 354 개 확인

bo_ts <- bostonData[-idx, ]
dim(bo_ts) # 506 개 * 0.3 = 152 개 확인
```

총 506 개 데이터 중 70%를 훈련용 데이터로, 나머지 30%를 검증용 데이터로 샘플링 하였다.

변수에 따른 주택가격의 변화를 확인하기 위해서는 모든 변수를 포함한 분석값이 존재해야 한다. lm 함수를 사용해 다중회귀분석을 실시하였다.

```
result.lm <- lm(formula = medv~., data = bo_tr)
result.lm
summary(result.lm)
```

위의 코드에 대한 결과 값은 아래와 동일하게 출력된다.

```
> summary(result.lm)
Call :
lm(formula = medv ~ ., data = bo_tr)
Residuals:
```

Min	1Q	Median	3Q	Max
-15.595	-2.730	-0.518	1.777	26.199

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.46	5.103	7.144	3.28e-12
Crim	-0.108	0.03286	-3.287	0.001087
Zn	0.04642	0.01373	3.382	0.000778
indus	0.02056	0.0615	0.334	**0.738288**
Chas	2.687	0.8616	3.118	0.001925
Nox	-17.77	3.82	-4.651	4.25e-06
Rm	3.81	0.4179	9.116	< 2e-16
Age	0.0006922	0.01321	0.052	**0.958229**
Dis	-1.476	0.1955	-7.398	6.01e-13
Rad	0.306	0.06635	4.613	5.07e-06
Tax	-0.01233	0.00376	-3.280	0.001112
Ptratio	-0.9527	0.1308	-7.283	1.31e-12
B	0.009312	0.002686	3.47	0.000573
Lstat	-0.5248	0.05072	-10.347	< 2e-16

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

해당 결과에서 변수들의 p-value 값을 보았을 때 유의수준 0.05 보다 훨씬 큰 변수들은 유의하지 않다고 볼 수 있기 때문에 **후진 제거법**을 실시해 영향이 적은 변수들을 제거할 것이다..

```
# 변수 선택(후진 제거법) - 변수가 많을수록 다른 선택법 보다 시간적으로 효율적
step <- step(result.lm, direction = 'backward')
formula(step)
```

위의 코드에 대한 결과 값은 아래와 동일하게 출력된다.

	Df	Sum of Sq	RSS	AIC
Age	1	0.06	11079	**1587.7**
Indus	1	2.52	11081	**1587.8**
<none>			11079	1589.6
Chas	1	218.97	11298	1597.5
Tax	1	242.26	11321	1598.6
Crim	1	243.22	11322	1598.6
Zn	1	257.49	11336	1599.3
B	1	270.63	11349	1599.8
Rad	1	479.15	11558	1609.1
Nox	1	487.16	11566	1609.4
Ptratio	1	1194.23	12273	1639.4
Dis	1	1232.41	12311	1641.0
Rm	1	1871.32	12950	1666.6
lstat	1	2410.84	13490	1687.3

```
# lm 함수 결과
Call :
lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
                                tax + ptratio + b + lstat, data = bostonData)
```

해당 결과를 통해 AIC 가 가장 낮은 변수는 Age, Indus 이 두 변수임을 알 수 있고, lm 함수 결과로부터 Age, Indus 두개의 변수가 빠진 것을 확인 할 수 있다.

Age, Indus 변수 제거 후 다시 다중회귀분석을 실시하였다.

```
result2.lm <- lm(formula = step, data = bo_tr)
result2.lm
summary(result2.lm)
```

위의 코드에 대한 결과 값은 아래와 동일하게 출력된다

```
> summary(result2.lm)
Call :
lm(formula = step, data = bo_tr)
Residuals:
```

Min	1Q	Median	3Q	Max
-15.9146	-2.6076	0.6624	1.7950	26.6233

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26.964605	6.317852	4.268	2.56e-05
Crim	-0.111504	0.035264	-3.162	0.001707
Zn	0.047037	0.015544	3.026	0.002665
Chas	2.712302	0.963911	2.814	0.005178
Nox	**-18.928319**	4.004869	-4.726	3.35e-06
Rm	**4.800240**	0.528360	9.085	< 2e-16
Dis	**-1.363232**	0.215102	-6.338	7.36e-10
Rad	0.261290	0.071417	3.659	0.000294
Tax	-0.011877	0.003781	-3.141	0.001829
Ptratio	**-0.851209**	0.152907	-5.567	5.25e-08
B	0.009343	0.003185	2.934	0.003573
Lstat	-0.361711	0.059422	-6.087	3.09e-09

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 4.651 on 342 degrees of freedom
Multiple R-squared:  0.7499,    Adjusted R-squared:  0.7419
F-statistic: 93.22 on 11 and 342 DF,  p-value: < 2.2e-16
```

해당 결과 Nox, Rm, Dis, Ptratio 4 개 변수의 Estimate 가 큰 것으로 보아 높은 상관성이 있다고 볼 수 있다.

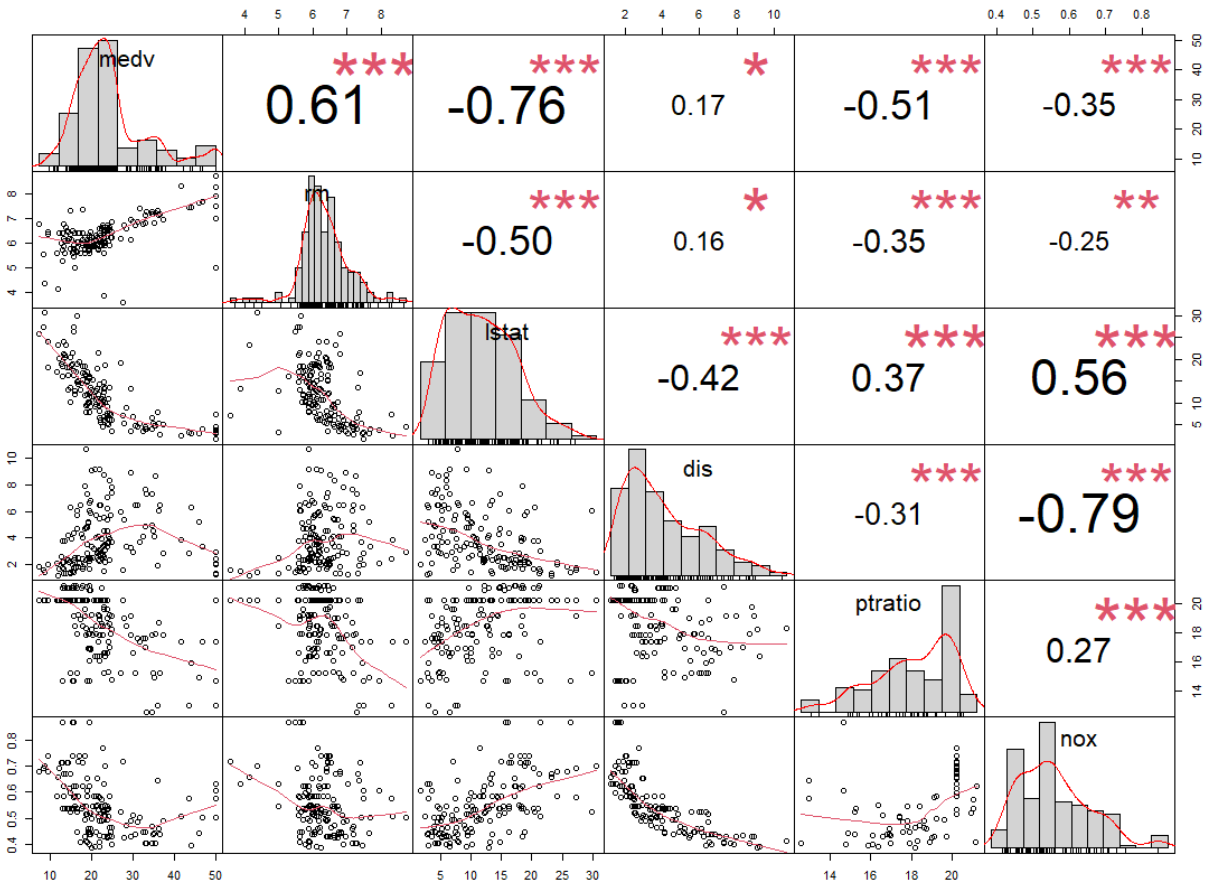
다음으로 예측 평가를 실시 하였다.

```
pred <- predict(result2.lm, bo_ts)
cor(pred, bo_ts$medv)  # 결과 : 0.8349967
```

predict 함수와 cor 함수를 사용하여 예측값과 실제값의 상관도를 구할 수 있다. 상관도는 0.8349967 로 높은 상관도가 있는 것을 알 수 있다.

다음으로 상관성이 높은 변수 위주로 시각화를 실시 하였다.

```
library(PerformanceAnalytics)
sd <- select_at(bo_ts, vars(medv,rm,lstat,dis,ptratio,nox))
chart.Correlation(sd,histogram=T,pch=20)
```



해당 결과의 상관계수를 보았을때 Rm(0.61), Lstat(-0.76), Ptratio(-0.51) 3 개의 변수가 높은 상관성이 있음을 알 수 있다. (상관 계수는 -1, +1 사이값으로 1 에 가까울수록 상관성이 높다)

마지막으로 다중회귀분석 모델 평가를 실시하였다.

```
install.packages("Metrics")
library(Metrics)
# RMSE : 평균 제곱근 오차 (낮을수록 오차가 적다.)
rmse(bo_ts$medv, pred)   # 결과 : 5.128862
```

해당 결과 RMSE 가 5.128862 임을 알 수 있다. 이 RMSE 수치를 뒤에서 랜덤포레스트 기법과 비교하여 어떤 예측 기법이 더 정확한지 비교 해 볼 예정이다.

2) 랜덤포레스트 기법을 사용한 데이터 분석

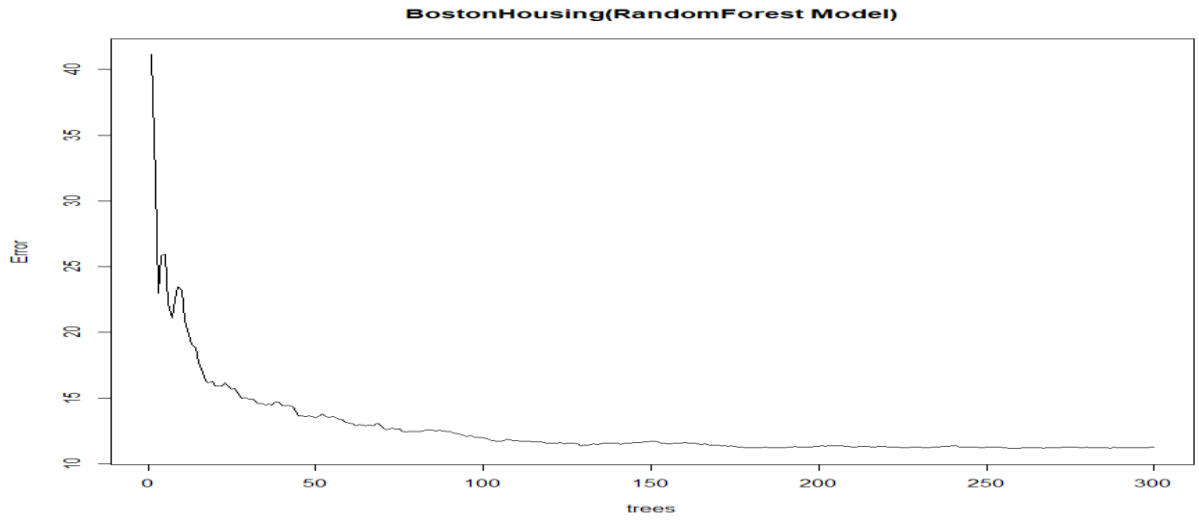
이어서 랜덤포레스트 기법을 통해 각각의 변수들이 Boston 주택가격에 어떠한 영향을 미치는지에 대하여 확인하고자 한다. 따라서 다중회귀분석과 동일하게 medv 변수를 종속변수로 설정하고 나머지 변수를 독립변수로 설정하였다.

랜덤포레스트 모델을 Tree 수 기준으로 100, 300, 50, 10 개짜리로 4 개의 모델을 생성하였다.

```
RFmodel100 <- randomForest(medv ~., data = bo_tr, ntree = 100, proximity=T)
RFmodel100

RFmodel300 <- randomForest(medv ~., data = bo_tr, ntree = 300, proximity=T)
RFmodel300

RFmodel50 <- randomForest(medv ~., data = bo_tr, ntree = 50, proximity=T)
RFmodel50

RFmodel10 <- randomForest(medv ~., data = bo_tr, ntree = 10, proximity=T)
RFmodel10
```

Tree수 기준으로 4개의 모델을 만든 이유는 랜덤포레스트 기법 자체에서 Tree수가 예측에 어떤 영향을 미치는지 추가적으로 알아보기 위해 생성하였다.

다음으로 Tree 수에 따른 Error 개수를 시각화 하였다.

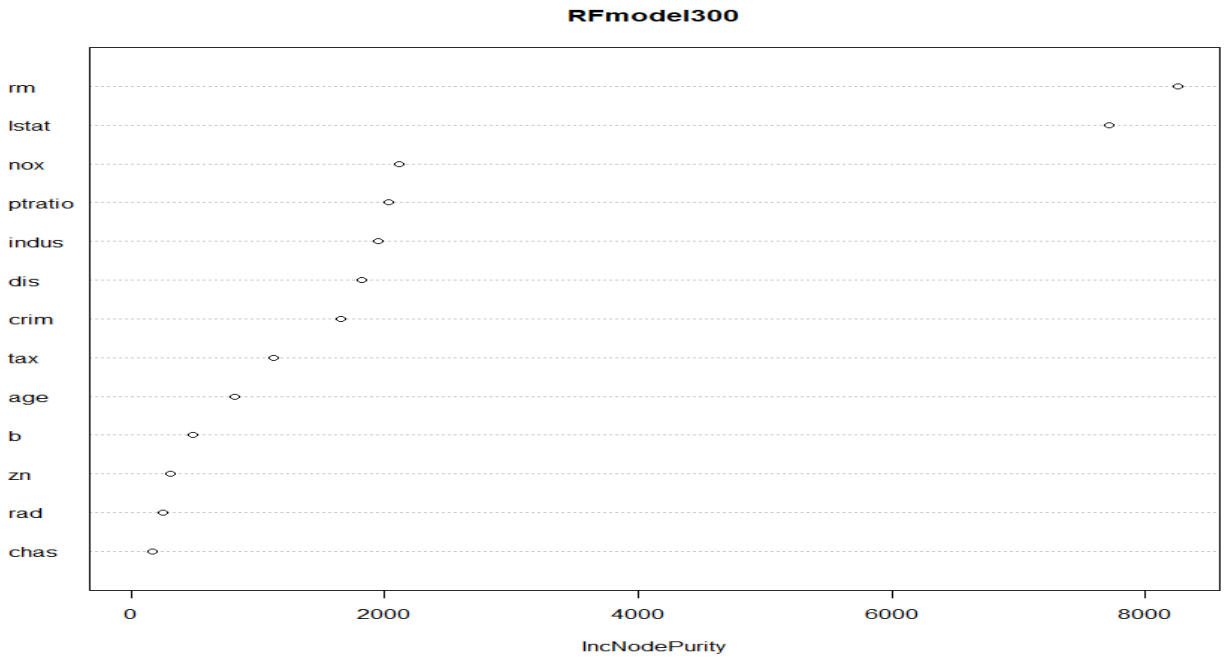


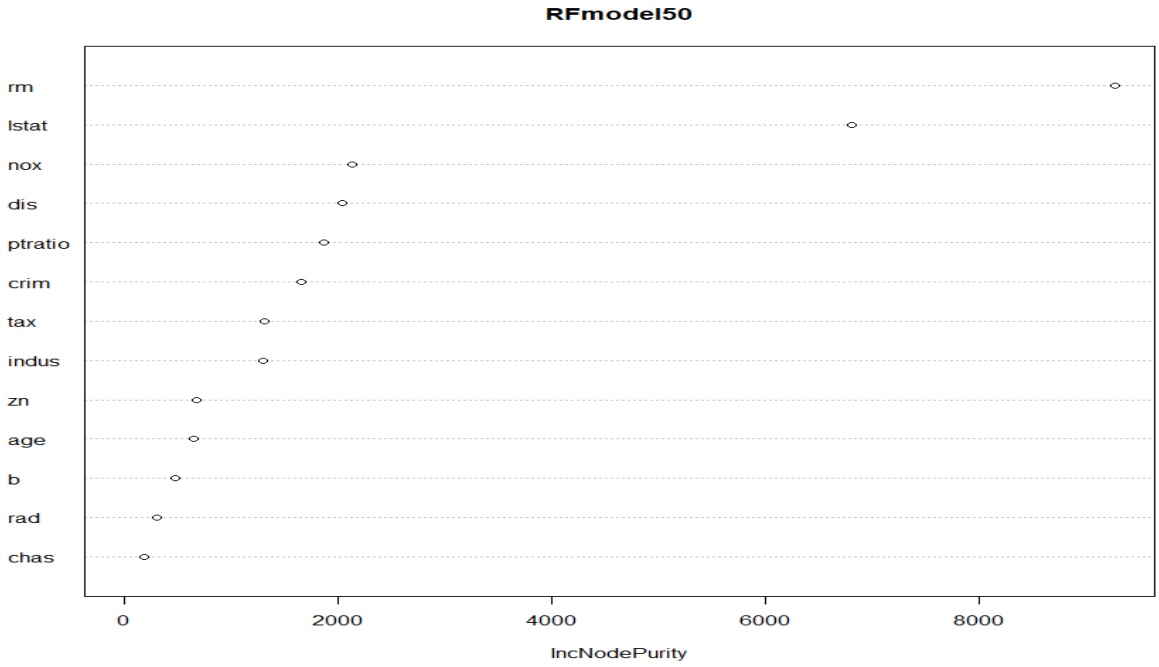
해당 결과 Tree 개수가 증가할수록 Error 가 현저히 감소하지만 일정 Tree 개수 이후로는 큰 의미가 없는 것을 알 수 있다.

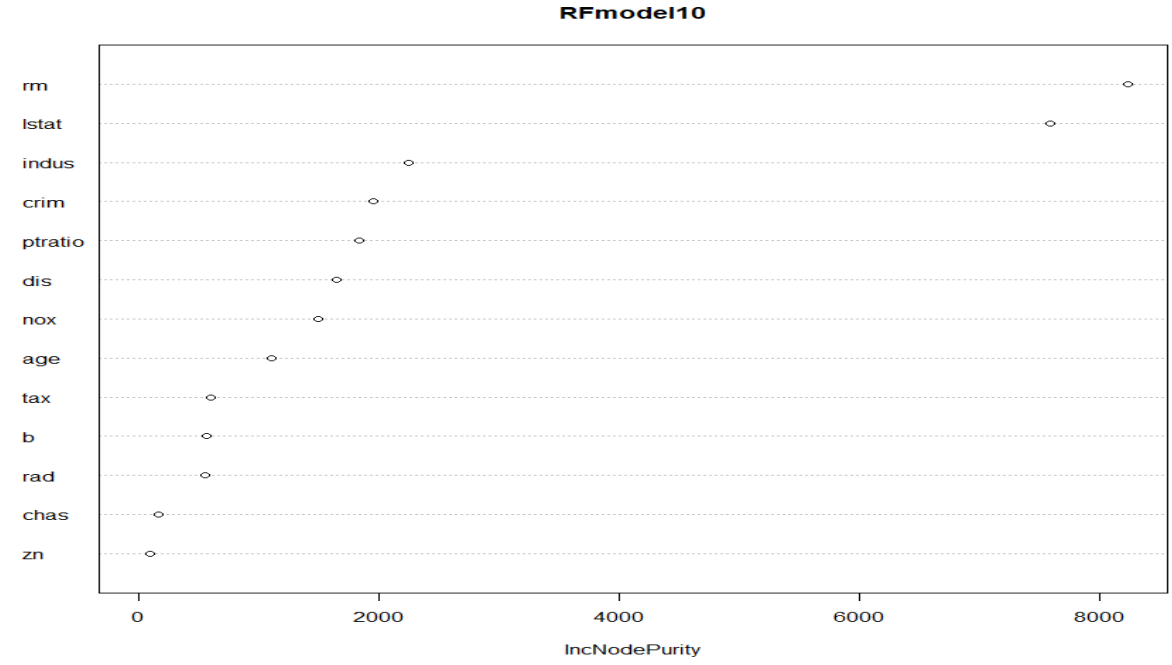
다음으로 importance 함수와 varImpPlot 함수로 중요 변수를 확인하고 시각화를 실시하였다.

```
importance(RFmodel300)
varImpPlot(RFmodel100)
```

	IncNodePurity(순수도)
Crim	1649.6699
Zn	309.1577
Indus	1953.7213
Chas	163.5356
Nox	2116.9145
Rm	**8259.7152**
Age	817.3789
Dis	1813.3859
Rad	247.6228
Tax	1119.0814
Ptratio	2026.1696
B	491.0627
Lstat	**7710.9408**







순서대로 Tree 가 300 개일 때, 50 개일 때, 10 개일 때에 해당하는 결과이다. 변수들의 순수도가 달라진 것을 알 수 있으며 Rm, Lstat 두 변수를 제외하고는 순서가 뒤바뀌는 것도 알 수 있다.

다음으로 예측 평가를 실시하였다.

```
pred100 <- predict(RFmodel100, newdata=bo_ts)
cor(pred100, bo_ts$medv) # 예측값과 실제값 상관도 : 0.94592

pred300 <- predict(RFmodel300, newdata=bo_ts)
cor(pred300, bo_ts$medv) # 예측값과 실제값 상관도 : 0.9457515

pred50 <- predict(RFmodel50, newdata=bo_ts)
cor(pred50, bo_ts$medv) # 예측값과 실제값 상관도 : 0.9479699

pred10 <- predict(RFmodel10, newdata=bo_ts)
cor(pred50, bo_ts$medv) # 예측값과 실제값 상관도 : 0.9419845
```

해당 결과 Tree 수가 달라져도 예측값과 실제값의 상관도에는 큰 변함이 없다.

다음으로 RMSE(평균 제곱근 오차)를 기준으로 모델 평가를 진행하였다.

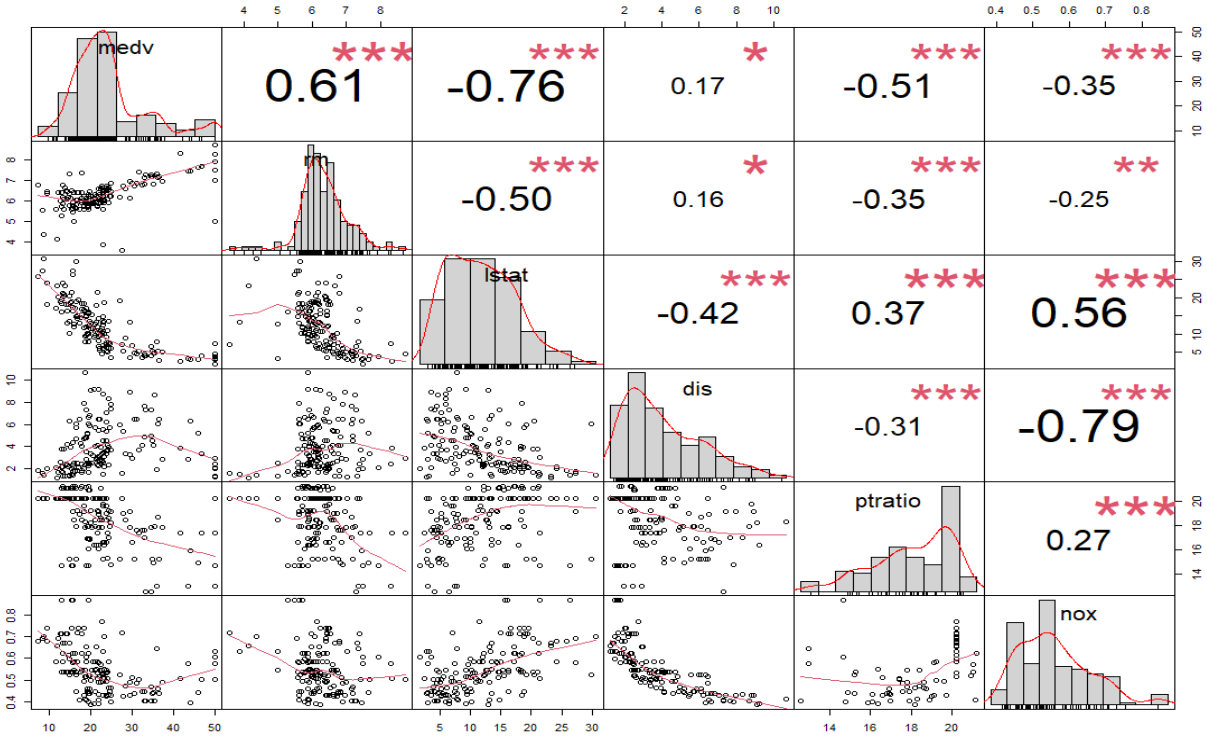
```
rmse(bo_ts$medv, pred100) # 3.213019
rmse(bo_ts$medv, pred300) # 3.225541
rmse(bo_ts$medv, pred50) # 3.125153
rmse(bo_ts$medv, pred10) # 3.476369
```

해당 결과 Tree 수를 100 개에서 300 개로 늘렸을 때 오차가 오히려 증가하였고 300 개에서 50 개로 줄였을 때 오차가 감소하였다. 하지만 50 개에서 10 개로 줄였을 때는 오히려 오차가 제일 심하게 나타나였다.

오차를 최소화 할 수 있는 적절한 Tree 수를 찾아 세팅하는 것이 관건이라고 볼 수 있다.

3. 분석 결과 및 결론

1) 데이터 분석 결과 및 결론 도출



	(Intercept)	Rm	Lstat	Dis	Ptratio	Nox
Estimate	26.964605	4.800240	-0.361711	-1.363232	-0.851209	-18.928319

시각화 자료와 회귀분석을 통한 변수들 계수 기준으로 Boston 주택 가격을 예측하였다.

변수의 영향이 없을때 Medv(주택 가격)은 26.97(x1000$)달러이다.

Rm(방의 개수)의 상관계수는 0.61, 변수계수는 4.8 로 주택 가격과 상관성이 높다는 것을 알 수 있고, 방의 개수가 1 개 증가할때마다 주택 가격이 4800 달러 증가 한다고 예측할 수 있다.

Lstat(하위계층 비율)의 상관계수는 -0.76, 변수계수는 -0.36 으로 주택 가격과 음의 상관성이 높다는 것을 알 수 있고, 하위계층 비율이 1% 증가 할때마다 주택 가격은 760 달러가 감소한다고 예측 할 수 있다.

Dis(직업센터 접근성 지수)의 상관계수는 0.17, 변수계수는 -1.36 으로 주택 가격과는 상관성이 거의 없다는 것을 알 수 있다.

Ptratio(학생/교사 비율)의 상관계수는 -0.51, 변수계수는 -0.85 로 주택 가격과 상관성이 있다는 것을 알 수 있고, 교사에 비해 학생의 수가 많을수록 주택 가격은 850 달러가 감소 함다고 예측 할 수 있다.

Nox(일산화질소 농도)의 상관계수는 -0.35, 변수계수는 -18.72 로 주택 가격과 상관성이 높지는 않지만 일산화질소의 농도가 높아질수록 주택 가격은 18720 달러가 감소한다고 예측 할 수 있기 때문에 아주 큰 폭으로 감소하는 것을 알 수 있다.

2) 다중회귀분석 / 랜덤포레스트 기법 비교

위에서 도출된 결과들로 볼 때 다중회귀분석과 랜덤포레스트 기법이 상관성이 높은 변수를 비슷하게 분류해 내었지만 완전 일치 하지는 않았다는 것을 알 수 있다.

다중선형회귀 분석 결과 상관도는 0.8349, RMSE(오차)는 5.12 로 도출되었다.

랜덤포레스트 분석 결과 상관도는 0.9458, RMSE(오차)는 3.21 로 도출되었다.

이로 미루어 봤을 때 BostonHousing 데이터셋에서는 다중회귀분석보다 랜덤포레스트 기법이 더 상관성이 높고 오차가 적어 향상된 분석 및 예측을 하였다고 도출할 수 있다.

참고 자료

1) 데이터셋 : mlbench 패키지 내 BostonHousing 데이터셋