Random Forest

Charles Hwang

12/17/2022

```
rm(list=ls())
sales<-read.csv("~/Desktop/Notes/Graduate/STAT 401 - Statistical Consulting/Fixed Data for Modeling.csv
train<-read.csv("~/Desktop/Notes/Graduate/STAT 401 - Statistical Consulting/Training Data.csv")[,-1]
test<-read.csv("~/Desktop/Notes/Graduate/STAT 401 - Statistical Consulting/Testing Data.csv")[,-1]
library(randomForest)
# All variables (36)
set.seed(1712)
rf<-randomForest(LastSalePrice~.-Valuation2019-Valuation2020-Valuation2021,data=train,importance=TRUE)
SRrf<-predict(rf,newdata=test)/test$LastSalePrice
ASRrf<-median(SRrf)          # 1.44693387522046 > 1.1
CODrf<-mean((SRrf-ASRrf)/ASRrf) # 14.9267116324929 < 15
data.frame(ASRrf,CODrf)
```

```
##      ASRrf     CODrf
## 1 1.446934 0.1492671
```

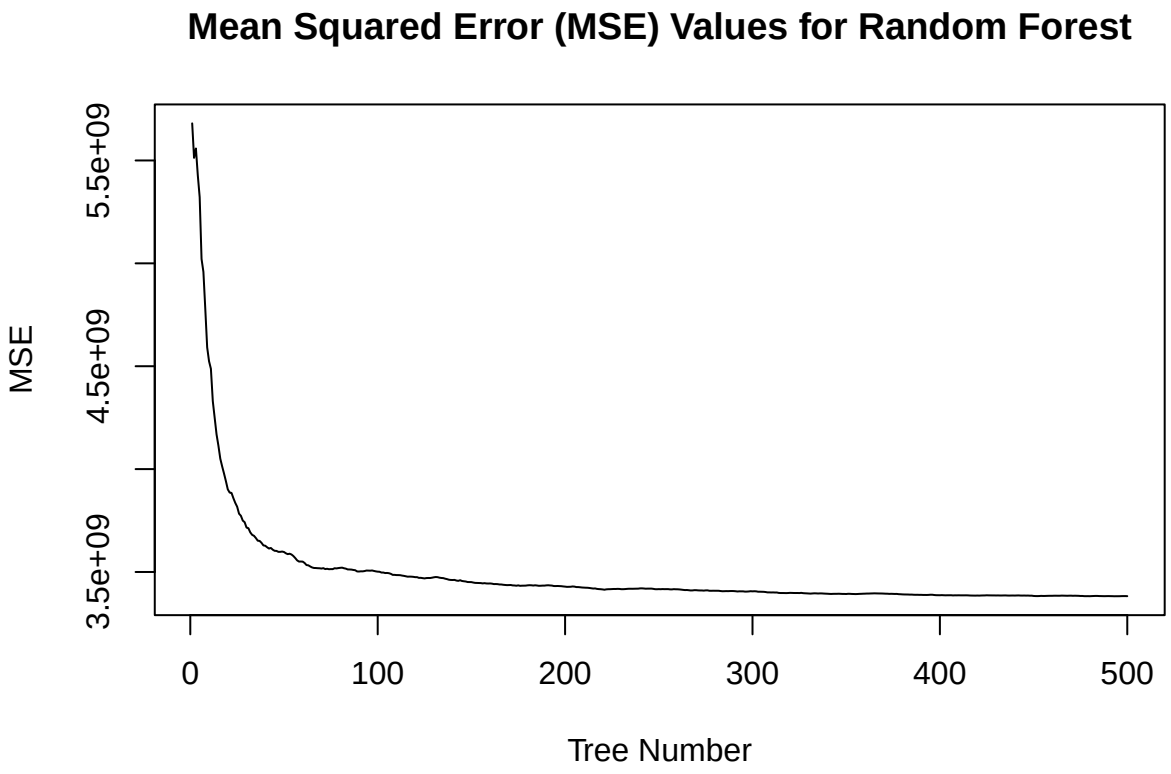
```
min(rf$mse) # MSE = 3,382,055,751
```

```
## [1] 3382055751
```

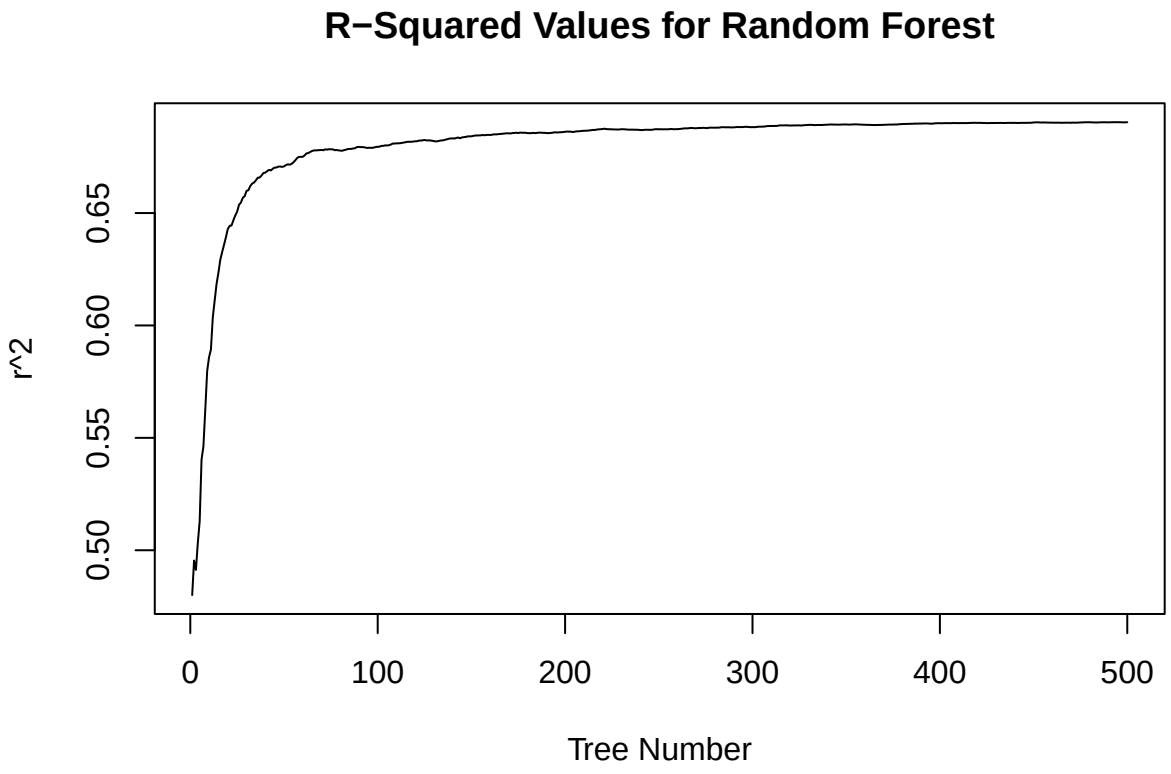
```
max(rf$rsq) # r^2 = 0.6904299
```

```
## [1] 0.6904299
```

```
plot(rf$mse,type="l",xlab="Tree Number",ylab="MSE",main="Mean Squared Error (MSE) Values for Random For
```

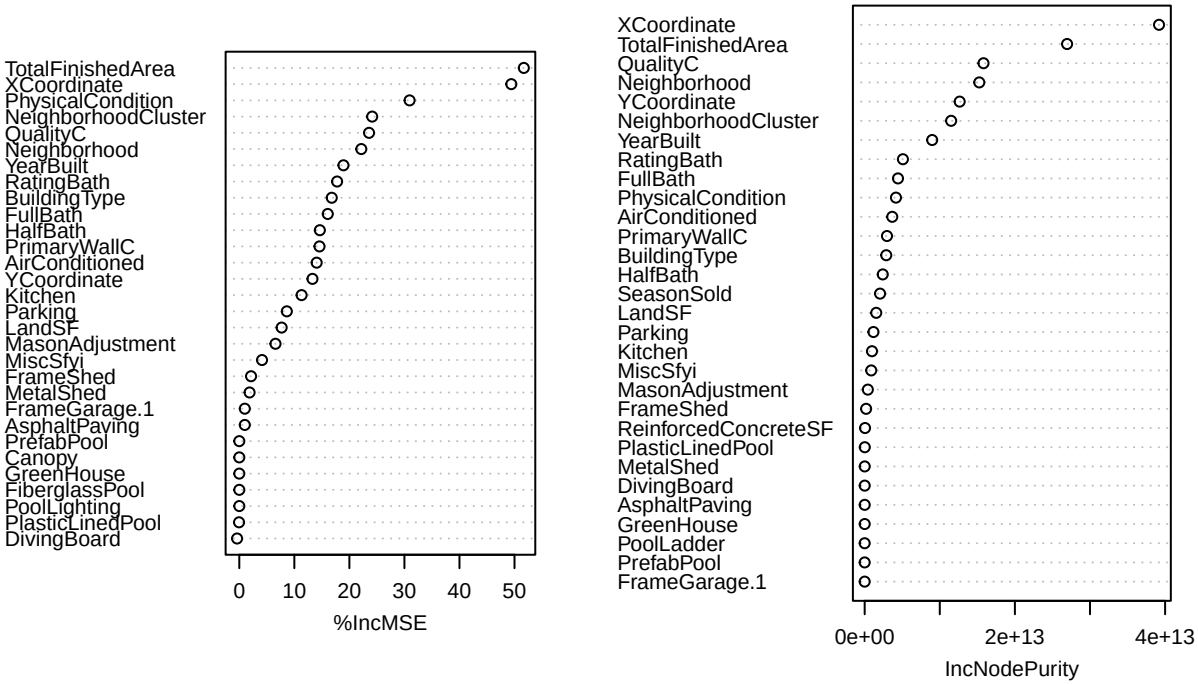



```
plot(rf$rsq,type="l",xlab="Tree Number",ylab="r^2",main="R-Squared Values for Random Forest")
```




```
varImpPlot(rf,main="Variable Importance Plot",cex=0.7)
```

Variable Importance Plot




```
# Variables from linear model (21)
set.seed(1712)
rflm<-randomForest(LastSalePrice~PhysicalCondition+Neighborhood+YearBuilt+FullBath+RatingBath+HalfBath+
SRlm<-predict(rflm,newdata=test)/test$LastSalePrice
ASRlm<-median(SRlm)              # 1.43075374414414 > 1.1
CODlm<-mean((SRlm-ASRlm)/ASRlm) # 16.6610997954028 > 15
data.frame(ASRlm,CODlm)
```

```
##      ASRlm    CODlm
## 1 1.430754 0.166611
```

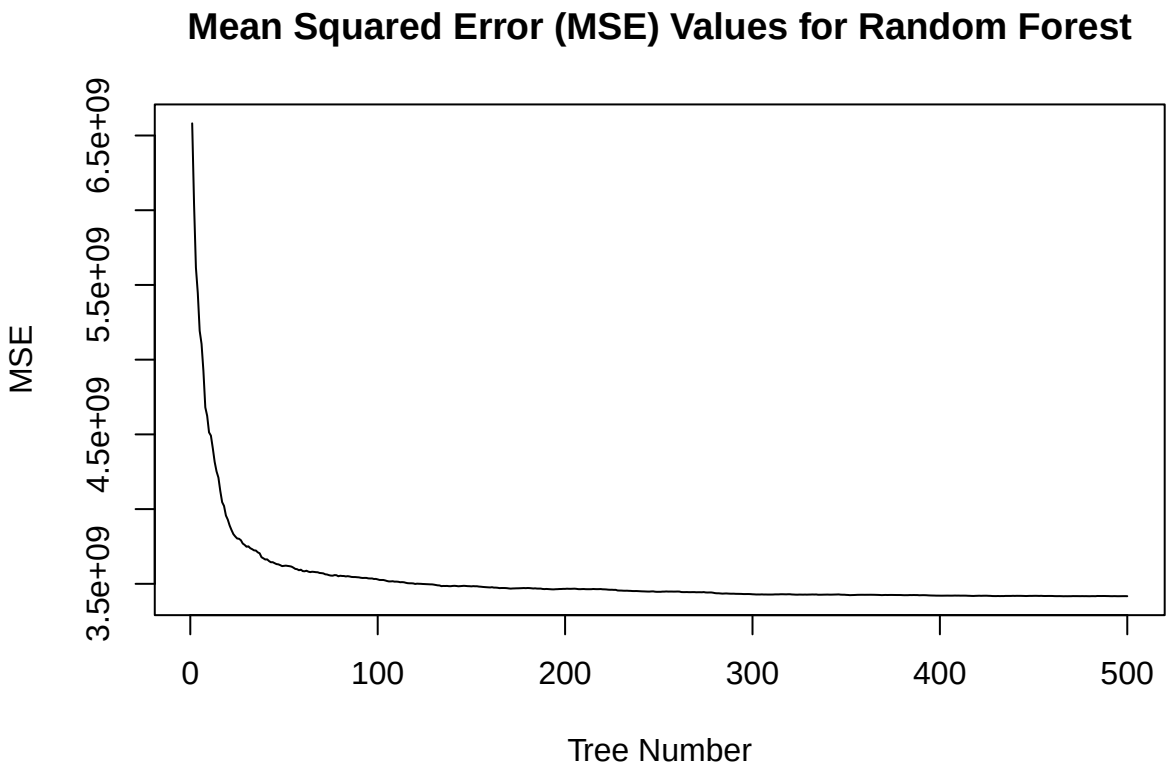
```
min(rflm$mse) # MSE = 3,416,858,940
```

```
## [1] 3416858940
```

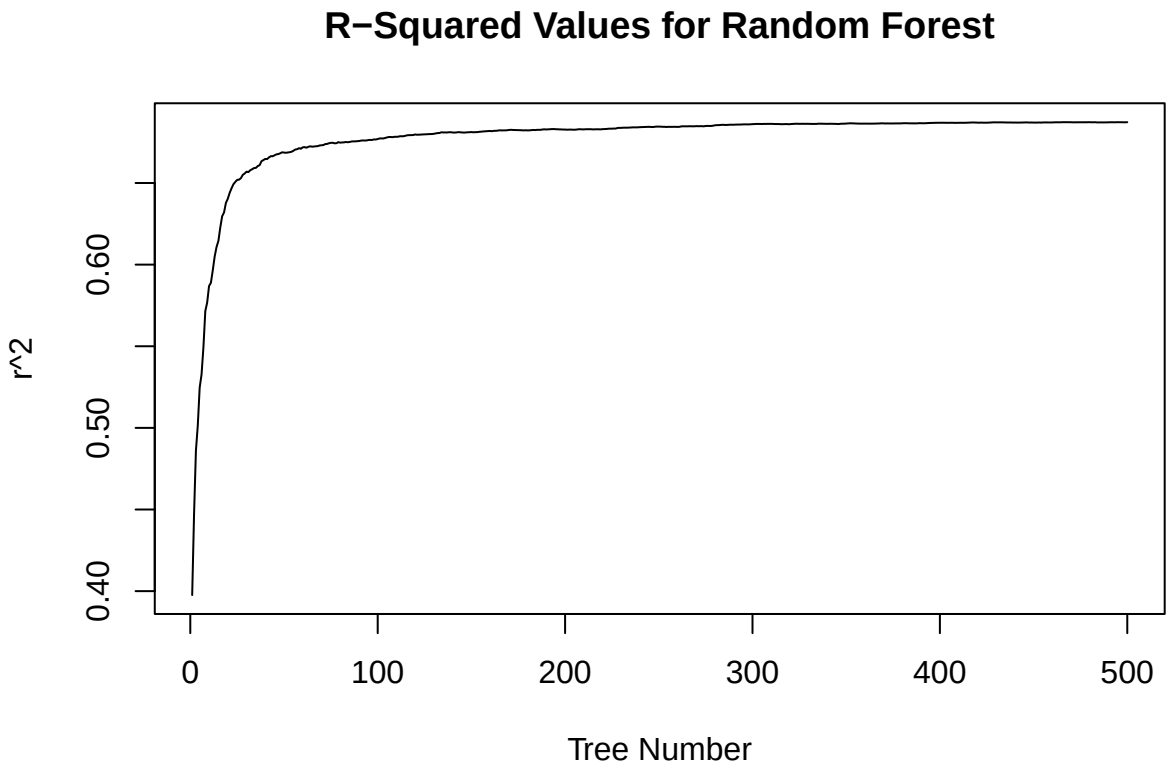
```
max(rflm$rsq) # r^2 = 0.6872443
```

```
## [1] 0.6872443
```

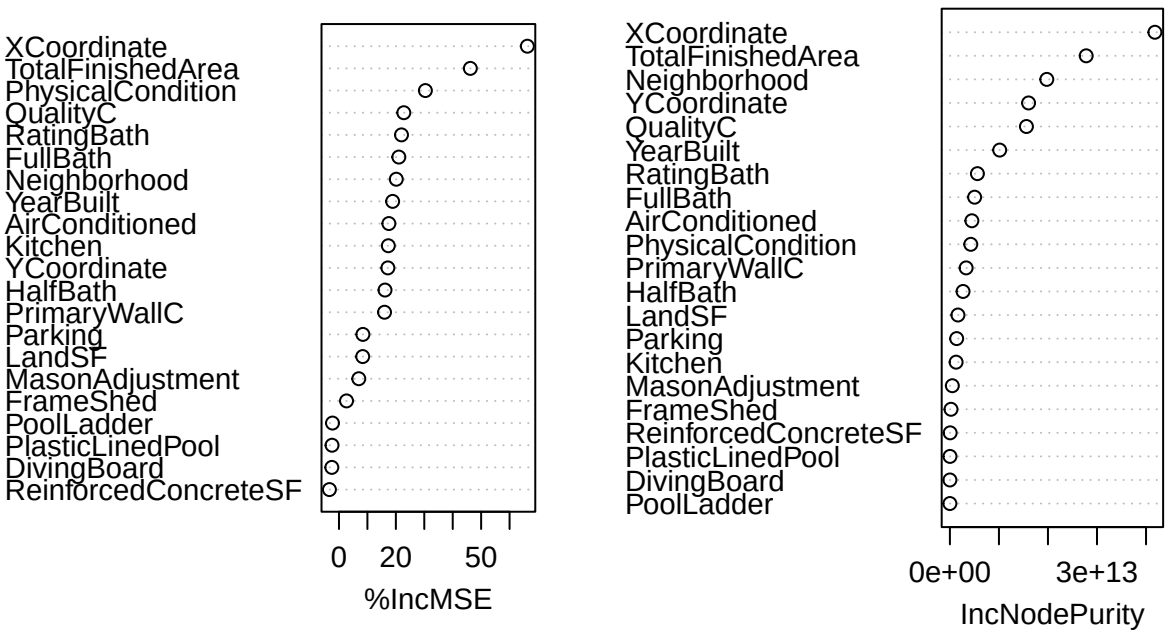
```
plot(rflm$mse,type="l",xlab="Tree Number",ylab="MSE",main="Mean Squared Error (MSE) Values for Random F
```




```
plot(rflm$rsq,type="l",xlab="Tree Number",ylab="r^2",main="R-Squared Values for Random Forest")
```




```
varImpPlot(rflm,main="Variable Importance Plot",cex=0.9)
```




```
# Variables from GLM created by bestglm except Quality (9)
set.seed(1712)
rfbestglm<-randomForest(LastSalePrice~PhysicalCondition+YearBuilt+HalfBath+TotalFinishedArea+LandSF+Air
SRbestglm<-predict(rfbestglm,newdata=test)/test$LastSalePrice
ASRbestglm<-median(SRbestglm)                    # 0.9 < 1.03893337961953 < 1.1
CODbestglm<-mean((SRbestglm-ASRbestglm)/ASRbestglm) # 10.1856495222623 < 15
data.frame(ASRbestglm,CODbestglm)
```

```
##   ASRbestglm CODbestglm
## 1   1.038933  0.1018565
```

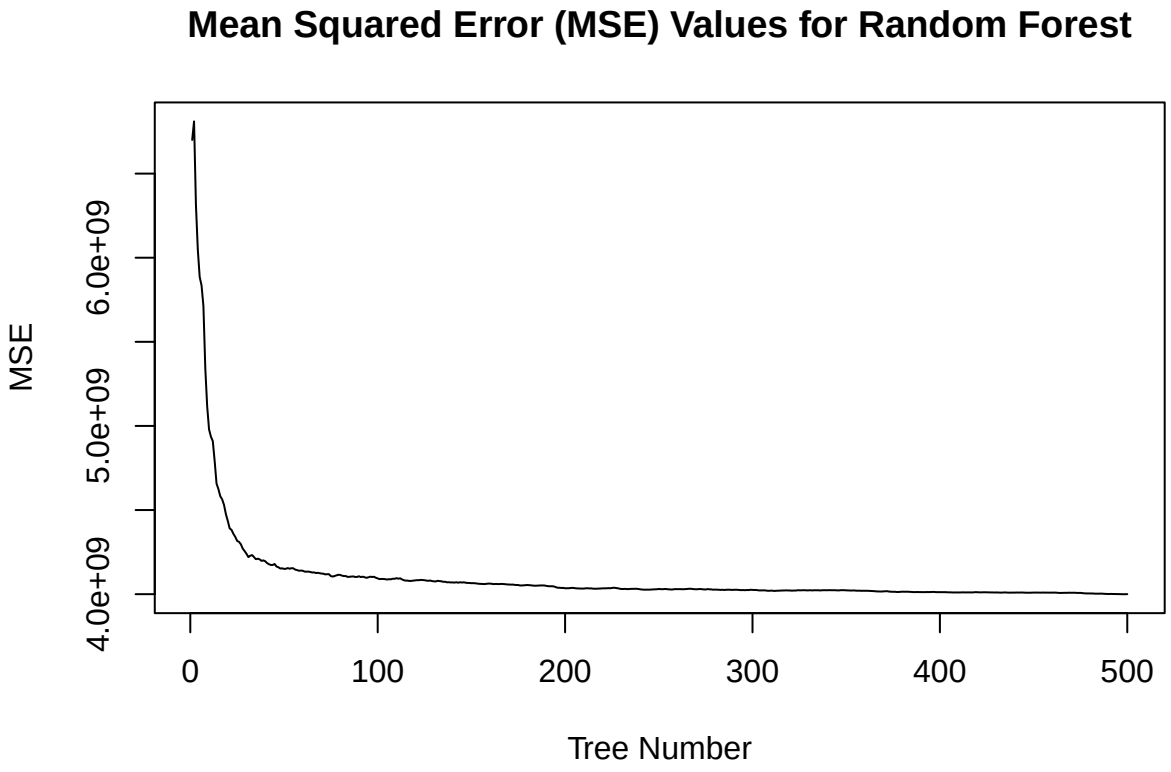
```
min(rfbestglm$mse) # MSE = 3,999,503,178
```

```
## [1] 3999503178
```

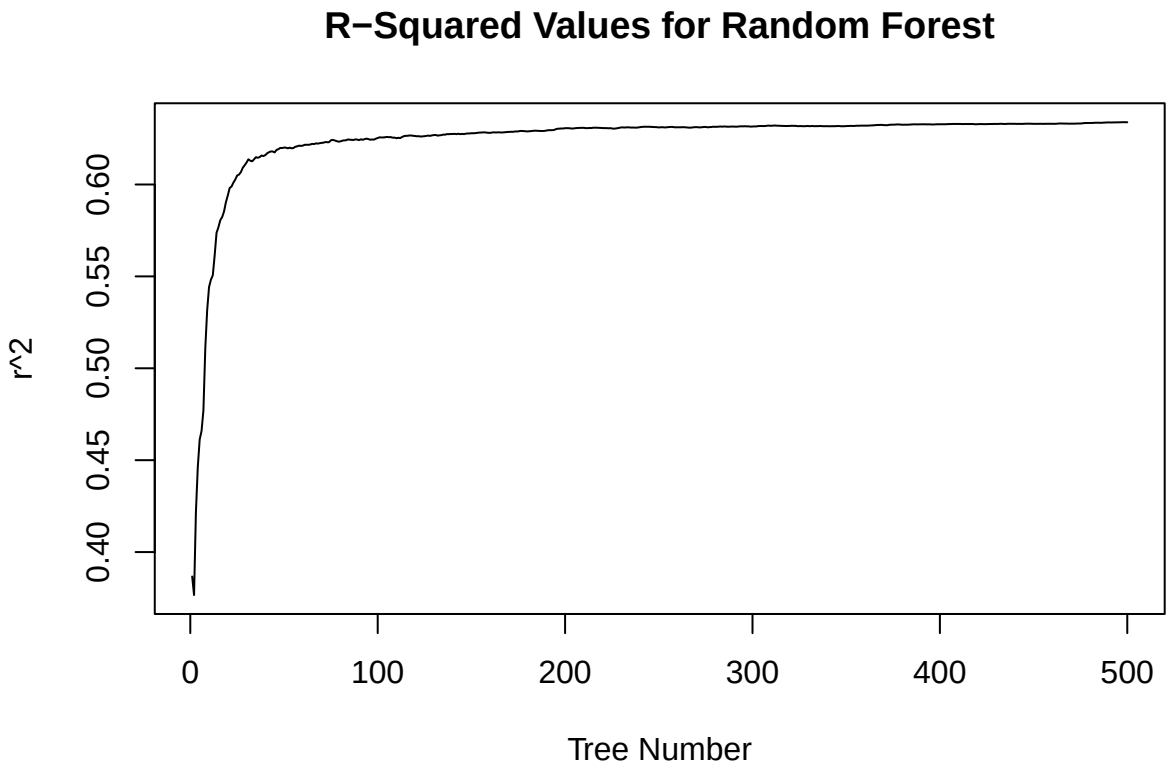
```
max(rfbestglm$rsq) # r^2 = 0.633913
```

```
## [1] 0.633913
```

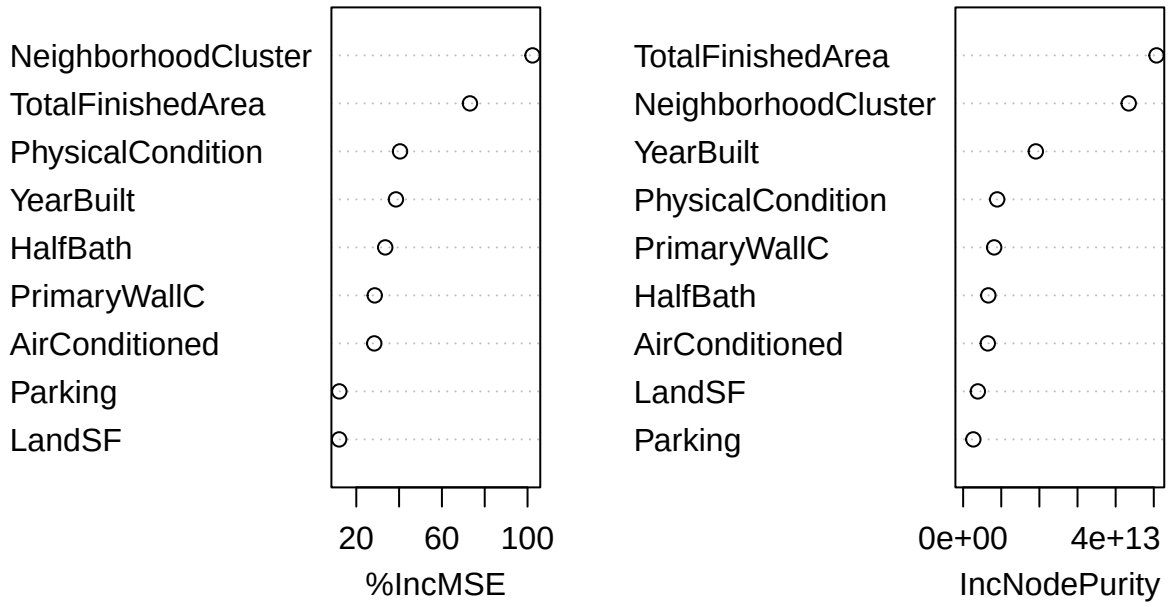
```
plot(rfbestglm$mse,type="l",xlab="Tree Number",ylab="MSE",main="Mean Squared Error (MSE) Values for Ran
```




```
plot(rfbestglm$rsq,type="l",xlab="Tree Number",ylab="r^2",main="R-Squared Values for Random Forest")
```




```
varImpPlot(rfbestglm,main="Variable Importance Plot")
```




```
# Variables from GLM created by purposeful variable selection (16)
set.seed(1712)
rfpvs<-randomForest(LastSalePrice~PhysicalCondition+YearBuilt+FullBath+RatingBath+HalfBath+Kitchen+Total
SRpvs<-predict(rfpvs,newdata=test)/test$LastSalePrice
ASRpvs<-median(SRpvs)                # 1.444461650770 89 > 1.1
CODpvs<-mean((SRpvs-ASRpvs)/ASRpvs) # 17.6974501019066 > 15
data.frame(ASRpvs,CODpvs)
```

```
##     ASRpvs    CODpvs
## 1 1.444462 0.1769745
```

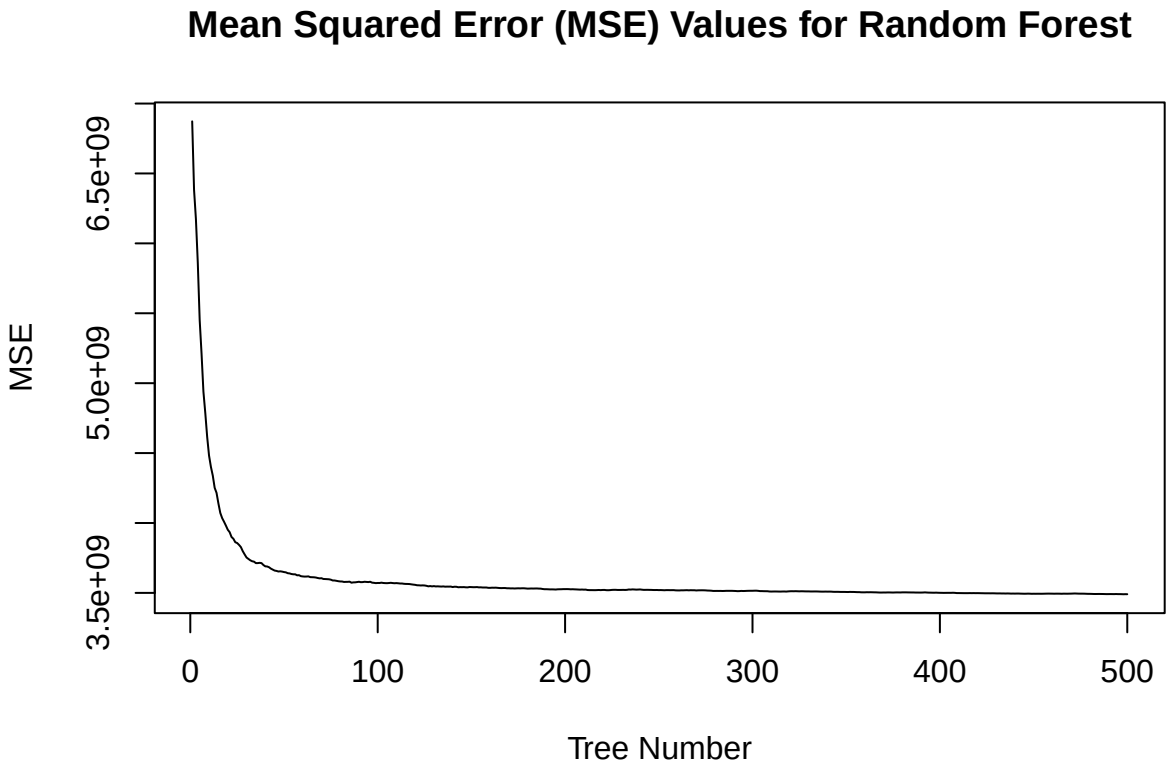
```
min(rfpvs$mse) # MSE = 3,490,417,322
```

```
## [1] 3490417322
```

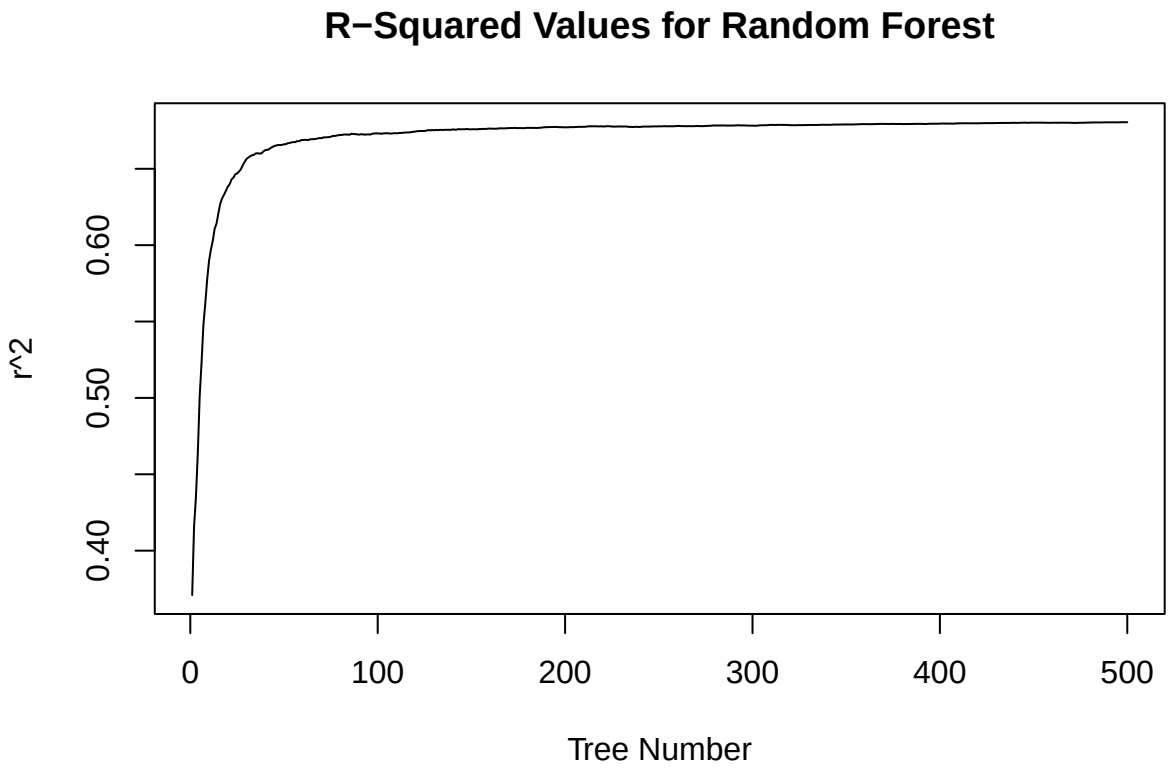
```
max(rfpvs$rsq) # r^2 = 0.6805112
```

```
## [1] 0.6805112
```

```
plot(rfpvs$mse,type="l",xlab="Tree Number",ylab="MSE",main="Mean Squared Error (MSE) Values for Random
```

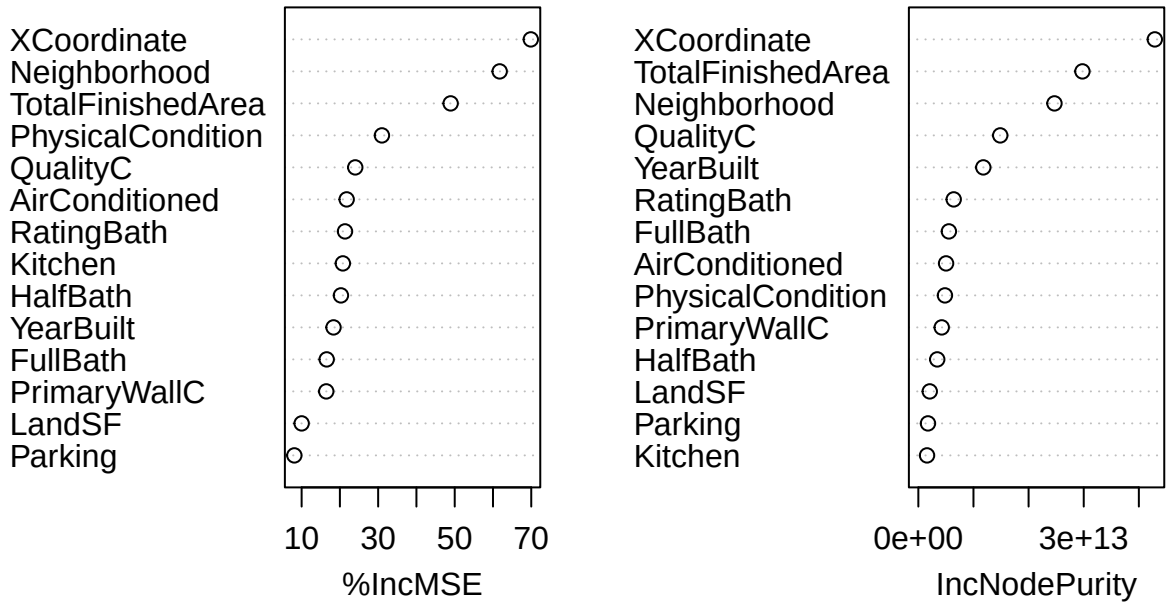



```
plot(rfpvs$rsq,type="l",xlab="Tree Number",ylab="r^2",main="R-Squared Values for Random Forest")
```




```
varImpPlot(rfpvs,main="Variable Importance Plot")
```

Variable Importance Plot




```
# Arbitrary choice of variables based on what intuitively made sense (12)
set.seed(1712)
rfchoice<-randomForest(LastSalePrice~BuildingType+PhysicalCondition+YearBuilt+FullBath+HalfBath+Kitchen+
SRchoice<-predict(rfchoice,newdata=test)/test$LastSalePrice
ASRchoice<-median(SRchoice)                # 0.9 < 1.03843279448145 < 1.1
CODchoice<-mean((SRchoice-ASRchoice)/ASRchoice) # 10.59617818973791 < 15
data.frame(ASRchoice,CODchoice)
```

```
##   ASRchoice CODchoice
## 1  1.038433 0.1059618
```

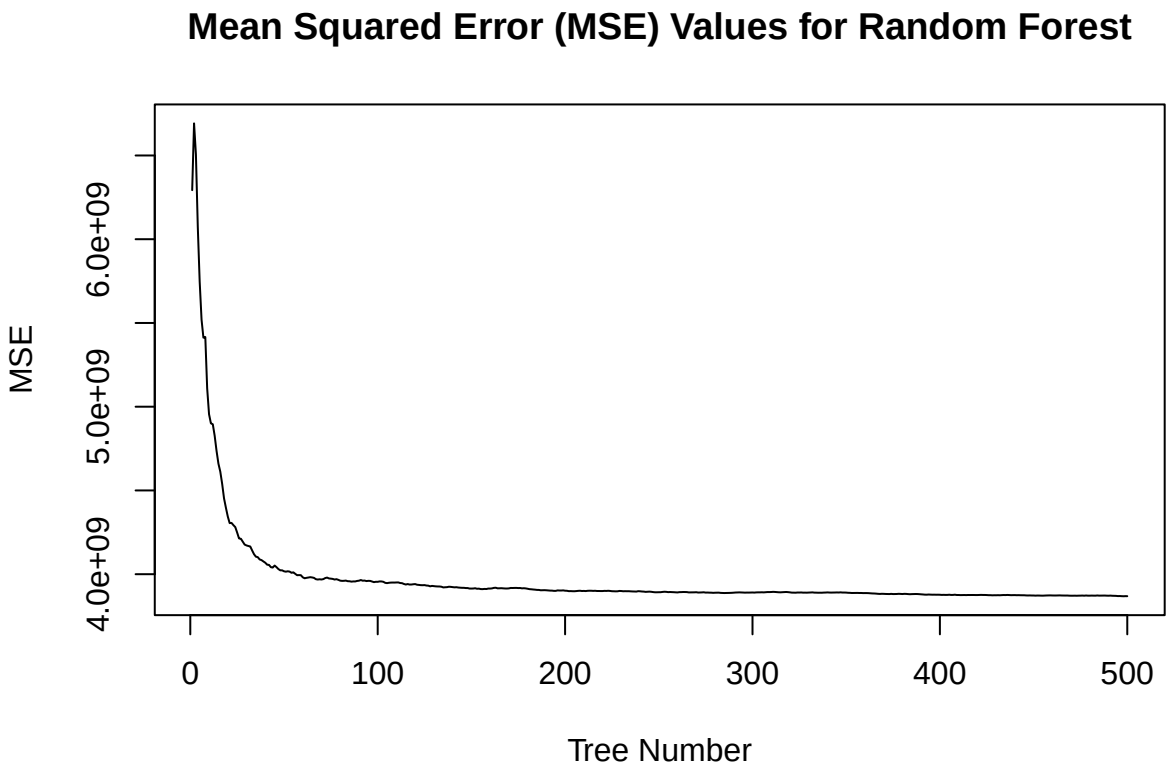
```
min(rfchoice$mse) # MSE = 3,868,303,043
```

```
## [1] 3868303043
```

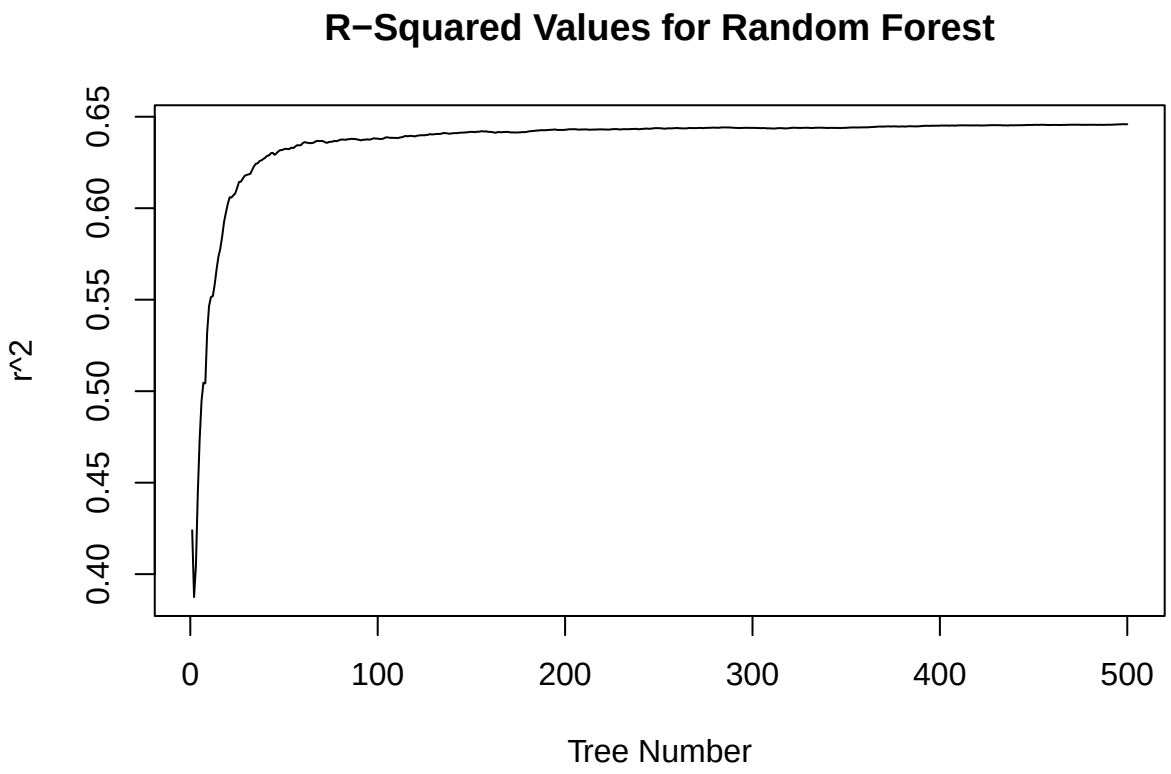
```
max(rfchoice$rsq) # r^2 = 0.6459222
```

```
## [1] 0.6459222
```

```
plot(rfchoice$mse,type="l",xlab="Tree Number",ylab="MSE",main="Mean Squared Error (MSE) Values for Rando
```




```
plot(rfchoice$rsq,type="l",xlab="Tree Number",ylab="r^2",main="R-Squared Values for Random Forest")
```




```
varImpPlot(rfchoice,main="Variable Importance Plot")
```

Variable Importance Plot