homework4_2

syh

May 31, 2017

```
# in this section we would like to train a simple tree first
# grid search for optimal combination of some parameters of rpart.control
```

```
# read all data (train + prediction) without missing value
real_all_data <- read.csv(file = "H:/kaggle/houseprice/data/real_all_data_hybrid.csv",
                          stringsAsFactors = FALSE)[,-c(1,2)]
```

```
# transform sale price to log sale price
real_all_data[,"SalePrice"] <- log(real_all_data[,"SalePrice"])
```

```
# convert categorical ones to factors
for(i in 1:dim(real_all_data)[2]){
  if(is.character(real_all_data[,i])){
    real_all_data[,i] <- as.factor(real_all_data[,i])
  }
}
```

```
# split all data into model and prediction
model_data <- real_all_data[1:1460,]
pre_x <- subset(real_all_data[-c(1:1460),], select = -SalePrice)
```

```
# split data into train and test
set.seed(1)
train_ind <- sample(1:dim(model_data)[1], size = dim(model_data)[1] * 0.7)
train_data <- model_data[train_ind,]
test_data <- model_data[-train_ind,]
```

```
# train a simple tree first
library(rpart)

forumla <- "SalePrice ~."
# all_control <- NULL
min_xerror <- Inf
opt_tree <- NULL

for(i in 1:30){

  seed.number = sample.int(10000, 1)[[1]]
  set.seed(seed.number)
  simple_tree <- rpart(formula = forumla, data = train_data, method = "anova",
                       control = rpart.control(
                          minsplit = sample(c(12, 21, 30, 39, 48), 1),
                          cp = sample(c(0.1, 0.01, 0),1)
                       )
                      )

  # all_control<- rbind(all_control, unlist(control))
  if(min(simple_tree$cptable[,"xerror"]) < min_xerror){
    opt_tree <- simple_tree
```

```
    min_xerror <- min(simple_tree$cptable[,"xerror"])
  }
}
```

```
opt_tree
```

```
## n= 1021
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##    1) root 1021 168.28670000 12.02963
##      2) OverallQual< 6.5 637  55.46297000 11.81480
##        4) Neighborhood=BrDale,BrkSide,Edwards,IDOTRR,MeadowV,OldTown 215  20.47921000 11.59573
##          8) GrLivArea< 1114.5 91   7.70568900 11.42802
##           16) CentralAir=N 23   2.77952100 11.14981 *
##           17) CentralAir=Y 68   2.54380000 11.52212
##             34) TotalBsmtSF< 664 28   0.86294350 11.40316 *
##             35) TotalBsmtSF>=664 40   1.00719500 11.60540
##               70) YearBuilt< 1938.5 19   0.52438210 11.51236 *
##               71) YearBuilt>=1938.5 21   0.16953850 11.68958 *
##          9) GrLivArea>=1114.5 124   8.33558000 11.71881
##           18) LotArea< 10415 101   5.01529300 11.65966
##             36) Exterior2nd=AsphShn,Brk Cmn,BrkFace,CBlock,Plywood 10   1.21258100 11.37472 *
##             37) Exterior2nd=AsbShng,CmentBd,HdBoard,MetalSd,Stucco,VinylSd,Wd Sdng,Wd Shng 91   2.90
##               74) TotalBsmtSF< 913.5 65   1.47340400 11.62667
##                148) GarageArea< 193.5 13   0.34507930 11.48143 *
##                149) GarageArea>=193.5 52   0.78552180 11.66298
##                  298) OverallCond< 5.5 18   0.30525070 11.57718 *
##                  299) OverallCond>=5.5 34   0.27760950 11.70840
##                    598) TotalBsmtSF>=782 12   0.13298460 11.64835 *
##                    599) TotalBsmtSF< 782 22   0.07774795 11.74116 *
##               75) TotalBsmtSF>=913.5 26   0.48760370 11.85172 *
##           19) LotArea>=10415 23   1.41490400 11.97858 *
##        5) Neighborhood=Blueste,ClearCr,CollgCr,Crawfor,Gilbert,Mitchel,NAmes,NPkVill,NridgHt,NWAmes,S
##         10) GrLivArea< 1151 162   4.62552600 11.76969
##           20) YearRemodAdd< 1950.5 13   0.48873540 11.47520 *
##           21) YearRemodAdd>=1950.5 149   2.91100900 11.79538
##             42) BsmtFullBath< 0.5 62   1.65944500 11.72516
##               84) GrLivArea< 951 31   0.68478800 11.64974
##                168) YearRemodAdd< 1971.5 21   0.49907890 11.60879 *
##                169) YearRemodAdd>=1971.5 10   0.07657922 11.73572 *
##               85) GrLivArea>=951 31   0.62198190 11.80058
##                170) OverallCond< 5.5 17   0.38554840 11.74577 *
##                171) OverallCond>=5.5 14   0.12333270 11.86714 *
##             43) BsmtFullBath>=0.5 87   0.72791270 11.84543
##               86) GarageArea< 345 32   0.18395180 11.78708
##                172) GarageArea< 267 11   0.06368985 11.73647 *
##                173) GarageArea>=267 21   0.07732635 11.81359 *
##               87) GarageArea>=345 55   0.37162700 11.87938
##                174) Fireplaces< 0.5 41   0.18305730 11.85630
##                  348) TotRmsAbvGrd< 5.5 29   0.08209465 11.83744 *
##                  349) TotRmsAbvGrd>=5.5 12   0.06573204 11.90187 *
##                175) Fireplaces>=0.5 14   0.10275220 11.94698 *
```

```
##         11) GrLivArea>=1151 260   8.32596400 12.02406
##           22) Neighborhood=Blueste,Mitchel,NAmes,NPkVill,Sawyer,SawyerW,SWISU 138   3.13666600 11.93
##             44) BsmtFinType1=None,Unf 28   0.81911800 11.81112 *
##             45) BsmtFinType1=ALQ,BLQ,GLQ,LwQ,Rec 110   1.76878900 11.96792
##               90) YearBuilt< 1963.5 60   0.82655320 11.91334
##                180) LotArea< 12331.5 50   0.50707570 11.88505
##                  360) YearRemodAdd< 1956 12   0.09309556 11.80162 *
##                  361) YearRemodAdd>=1956 38   0.30406760 11.91140
##                    722) LotArea< 7920.5 11   0.04663460 11.85310 *
##                    723) LotArea>=7920.5 27   0.20481760 11.93515 *
##                181) LotArea>=12331.5 10   0.07941511 12.05478 *
##               91) YearBuilt>=1963.5 50   0.54895560 12.03342
##                182) KitchenQual=Fa,TA 35   0.28875790 11.99711
##                  364) X1stFlrSF< 1066 11   0.04944272 11.93731 *
##                  365) X1stFlrSF>=1066 24   0.18193400 12.02453 *
##                183) KitchenQual=Gd 15   0.10639330 12.11814 *
##           23) Neighborhood=ClearCr,CollgCr,Crawfor,Gilbert,NridgHt,NWAmes,Somerst,Timber,Veenker 122
##             46) BsmtFinSF1< 690.5 88   1.42585100 12.07345
##               92) OverallQual< 5.5 15   0.21516220 11.95089 *
##               93) OverallQual>=5.5 73   0.93904220 12.09864
##                186) GrLivArea< 1695 53   0.32737880 12.06301
##                  372) WoodDeckSF< 191 43   0.18664230 12.04712
##                    744) GrLivArea< 1414 14   0.07926544 11.98749 *
##                    745) GrLivArea>=1414 29   0.03357055 12.07590 *
##                  373) WoodDeckSF>=191 10   0.08314647 12.13137 *
##                187) GrLivArea>=1695 20   0.36612660 12.19305 *
##             47) BsmtFinSF1>=690.5 34   0.69658880 12.25317
##               94) GrLivArea< 1639 16   0.16027890 12.14738 *
##               95) GrLivArea>=1639 18   0.19812980 12.34719 *
##       3) OverallQual>=6.5 384  34.65703000 12.38600
##         6) OverallQual< 7.5 213   8.10795300 12.22113
##          12) GrLivArea< 2027 177   5.33050100 12.17929
##            24) GarageArea< 407.5 31   0.49936430 11.97771
##              48) Neighborhood=Edwards,IDOTRR,NAmes,OldTown,SawyerW 14   0.12806490 11.86886 *
##              49) Neighborhood=Blmngtn,BrkSide,CollgCr,Crawfor,Gilbert,Timber 17   0.06885572 12.06734
##            25) GarageArea>=407.5 146   3.30398400 12.22209
##              50) TotalBsmtSF< 766 19   0.17549310 12.03025 *
##              51) TotalBsmtSF>=766 127   2.32465100 12.25079
##               102) Neighborhood=Blmngtn,BrkSide,CollgCr,Gilbert,Mitchel,NAmes,NridgHt,NWAmes,SawyerW,
##                 204) Exterior1st=CemntBd,HdBoard,Plywood,WdShing 15   0.08691940 12.08560 *
##                 205) Exterior1st=MetalSd,VinylSd,Wd Sdng 83   0.79351430 12.23833
##                   410) BsmtFinSF1< 728.5 65   0.49373140 12.21317
##                     820) LotArea< 9461.5 34   0.17243340 12.17915
##                      1640) BsmtUnfSF>=931 20   0.06948196 12.15250 *
##                      1641) BsmtUnfSF< 931 14   0.06845350 12.21722 *
##                     821) LotArea>=9461.5 31   0.23877990 12.25048
##                      1642) GarageArea< 505 12   0.05810060 12.19778 *
##                      1643) GarageArea>=505 19   0.12629090 12.28377 *
##                   411) BsmtFinSF1>=728.5 18   0.11007440 12.32918 *
##                103) Neighborhood=ClearCr,Crawfor,NoRidge,Somerst,StoneBr 29   0.59672270 12.37189 *
##          13) GrLivArea>=2027 36   0.94393180 12.42685
##            26) Neighborhood=Crawfor,Gilbert,Mitchel,NAmes,NWAmes,OldTown,SawyerW,StoneBr,SWISU 26   0
##            27) Neighborhood=ClearCr,NoRidge,NridgHt,Timber,Veenker 10   0.16974050 12.58452 *
##         7) OverallQual>=7.5 171  13.54744000 12.59136
```

```
##         14) OverallQual< 8.5 125   6.20940200 12.49818
##           28) X1stFlrSF< 995 14   0.60540550 12.18334 *
##           29) X1stFlrSF>=995 111   4.04122900 12.53789
##             58) GrLivArea< 2044 69   2.00953400 12.46158
##              116) BsmtFinSF1< 940.5 40   0.93447480 12.38167
##                232) LotArea< 11072.5 25   0.50242650 12.32909 *
##                233) LotArea>=11072.5 15   0.24777500 12.46929 *
##              117) BsmtFinSF1>=940.5 29   0.46732260 12.57180 *
##             59) GrLivArea>=2044 42   0.96970210 12.66326
##              118) X1stFlrSF< 1383.5 23   0.35589810 12.58506 *
##              119) X1stFlrSF>=1383.5 19   0.30286470 12.75793 *
##         15) OverallQual>=8.5 46   3.30341200 12.84457
##           30) Neighborhood=CollgCr,Edwards,OldTown,Somerst,Timber 12   0.37407320 12.60691 *
##           31) Neighborhood=Gilbert,NoRidge,NridgHt,StoneBr,Veenker 34   2.01233600 12.92845
##             62) TotalBsmtSF< 1986 24   0.65687910 12.82945 *
##             63) TotalBsmtSF>=1986 10   0.55568050 13.16606 *
```

```
printcp(opt_tree)
```

```
##
## Regression tree:
## rpart(formula = forumla, data = train_data, method = "anova",
##     control = rpart.control(minsplit = sample(c(12, 21, 30, 39,
##         48), 1), cp = sample(c(0.1, 0.01, 0), 1)))
##
## Variables actually used in tree construction:
##  [1] BsmtFinSF1   BsmtFinType1 BsmtFullBath BsmtUnfSF    CentralAir
##  [6] Exterior1st  Exterior2nd  Fireplaces   GarageArea   GrLivArea
## [11] KitchenQual  LotArea      Neighborhood OverallCond  OverallQual
## [16] TotalBsmtSF  TotRmsAbvGrd WoodDeckSF   X1stFlrSF    YearBuilt
## [21] YearRemodAdd
##
## Root node error: 168.29/1021 = 0.16483
##
## n= 1021
##
##            CP nsplit rel error  xerror     xstd
## 1  0.46448526      0   1.00000 1.00311 0.053368
## 2  0.09254674      1   0.53551 0.53747 0.032180
## 3  0.07725888      2   0.44297 0.45023 0.027006
## 4  0.03837436      3   0.36571 0.37332 0.024012
## 5  0.02637129      4   0.32733 0.33046 0.022931
## 6  0.02397471      5   0.30096 0.32546 0.022265
## 7  0.01415660      6   0.27699 0.28400 0.017744
## 8  0.01351739      7   0.26283 0.27871 0.017211
## 9  0.01132224      8   0.24931 0.27322 0.017071
## 10 0.01089522      9   0.23799 0.27192 0.016984
## 11 0.00928634     10   0.22710 0.26771 0.016521
## 12 0.00907471     11   0.21781 0.26365 0.016407
## 13 0.00728389     12   0.20874 0.25670 0.016277
## 14 0.00631062     13   0.20145 0.25820 0.017037
## 15 0.00547193     14   0.19514 0.25423 0.017022
## 16 0.00544905     16   0.18420 0.24868 0.016742
## 17 0.00477661     17   0.17875 0.24693 0.016159
## 18 0.00475246     18   0.17397 0.24365 0.016070
```

```
## 19 0.00470662     19   0.16922 0.24365 0.016070
## 20 0.00400306     20   0.16451 0.24105 0.016035
## 21 0.00361132     21   0.16051 0.23588 0.015523
## 22 0.00327517     22   0.15690 0.23330 0.015440
## 23 0.00326086     23   0.15362 0.23086 0.015347
## 24 0.00311166     24   0.15036 0.22972 0.015345
## 25 0.00233697     25   0.14725 0.22321 0.015104
## 26 0.00209568     26   0.14491 0.21901 0.015010
## 27 0.00204522     27   0.14282 0.22056 0.015085
## 28 0.00203701     28   0.14077 0.22090 0.015087
## 29 0.00200955     29   0.13874 0.22108 0.015085
## 30 0.00186155     30   0.13673 0.22139 0.015064
## 31 0.00184768     31   0.13487 0.21964 0.014992
## 32 0.00179719     32   0.13302 0.21964 0.014992
## 33 0.00176085     33   0.13122 0.21908 0.014983
## 34 0.00161419     34   0.12946 0.21783 0.014970
## 35 0.00145904     35   0.12785 0.21706 0.014950
## 36 0.00142651     36   0.12639 0.21589 0.015138
## 37 0.00120426     37   0.12496 0.21271 0.015094
## 38 0.00112729     38   0.12376 0.20978 0.014911
## 39 0.00109500     39   0.12263 0.21004 0.014935
## 40 0.00102405     40   0.12153 0.20932 0.014908
## 41 0.00091394     41   0.12051 0.20908 0.014907
## 42 0.00067207     42   0.11960 0.20825 0.014938
## 43 0.00065313     43   0.11892 0.20870 0.014911
## 44 0.00064848     44   0.11827 0.20873 0.014911
## 45 0.00050995     45   0.11762 0.20882 0.014917
## 46 0.00049034     46   0.11711 0.20842 0.014922
## 47 0.00039740     47   0.11662 0.20809 0.014924
## 48 0.00039039     48   0.11622 0.20828 0.014923
## 49 0.00034097     50   0.11544 0.20833 0.014922
## 50 0.00032319     51   0.11510 0.20812 0.014925
## 51 0.00031265     52   0.11478 0.20811 0.014925
## 52 0.00025513     53   0.11447 0.20782 0.014927
## 53 0.00020935     54   0.11421 0.20795 0.014926
## 54 0.00020500     55   0.11400 0.20775 0.014929
## 55 0.00000000     56   0.11380 0.20762 0.014930
```

```
# find optimal cp
cptable <- as.data.frame(opt_tree$cptable)
opt_cp <- cptable[with(cptable, min(which(xerror - xstd <= min(xerror)))), "CP"]
```

```
# create a optimal tree with opt_cp
opt_tree <- prune(tree = opt_tree, cp = opt_cp)
```

```
est_test_sale_price <- predict(object = opt_tree, newdata = test_data)
```

```
# sse
sum((est_test_sale_price - test_data$SalePrice) ^ 2)
```

```
## [1] 17.62572
```

```
# let make a prediction
pre_sale_price <- predict(object = opt_tree, newdata = pre_x)
#
```

```r
result <- data.frame(Id = c(1461:2919), SalePrice = exp(pre_sale_price))

write.csv(x = result, file = "H:/kaggle/houseprice/data/submission_4.csv",
          row.names = FALSE)
# Your submission scored 0.21641
# even linear regression outperforms it!!!
```