University of Waterloo

STAT 444

STAT 444 Spring 2019

Group Gengyao_Yuan:

Gengyao Yuan(20613017)

Haohan Li(20610397)

Contents

1 Executive summary:

sample

sample
sample

2 Itroduction:

sample

3 Data

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

data$KITCHENS[is.na(data$KITCHENS)] <- getmode(data$KITCHENS[!is.na(data$KITCHENS)])



med_diff_built_remodel <-floor(median(data$YR_RMDL[!is.na(data$YR_RMDL)&!is.na(data$AYB)]-
                                data$AYB[!is.na(data$YR_RMDL)&!is.na(data$AYB)] ))


# Fill AYB and YR_RMDL
data$YR_RMDL[is.na(data$YR_RMDL)&!is.na(data$AYB)] <-data$AYB[is.na(data$YR_RMDL)&!is.na(data$AYB)]


#Missing Both
missing_both <- is.na(data$YR_RMDL)&is.na(data$AYB)

data$AYB[missing_both]<-floor(median(data$AYB[!is.na(data$AYB)]))

data$YR_RMDL[missing_both]<-data$AYB[missing_both]- med_diff_built_remodel


missing_built_have_remodel <- (!is.na(data$YR_RMDL)&is.na(data$AYB))
data$AYB[missing_built_have_remodel] <- data$YR_RMDL[missing_built_have_remodel]-med_diff_built_remodel

data$STORIES[is.na(data$STORIES)]<-floor(median(data$STORIES[!is.na(data$STORIES)]))
```

##Outliers / Extreme value

By eyeball the data, we can see there is several outliers

```
data[data$YR_RMDL==20,]
```

```
##          Id BATHRM HF_BATHRM       HEAT AC ROOMS BEDRM  AYB YR_RMDL  EYB
## 21997 21997      1         1 Forced Air  Y     8     4 1929      20 1967
##       STORIES            SALEDATE  PRICE  GBA   STYLE        STRUCT
## 21997       2 2015-08-06 00:00:00 335000 1640 2 Story Semi-Detached
##               GRADE CNDTN      EXTWALL     ROOF  INTWALL KITCHENS
```

```
## 21997 Above Average  Good Common Brick Built Up Hardwood        1
##       FIREPLACES USECODE LANDAREA         FULLADDRESS ZIPCODE
## 21997          1      13     2380 617 ONEIDA PLACE NW   20011
##             NATIONALGRID LATITUDE LONGITUDE ASSESSMENT_NBHD
## 21997 18S UJ 24811 14526  38.9622 -77.02199      Brightwood
##       ASSESSMENT_SUBNBHD CENSUS_TRACT CENSUS_BLOCK   WARD QUADRANT fold
## 21997   006 E Brightwood         1901  001901 1000 Ward 4       NW    2
```

```
data$YR_RMDL[data$YR_RMDL==20]=data$AYB[data$YR_RMDL==20]+med_diff_built_remodel
data$AC[data$AC ==0] <- getmode(data$AC[data$AC !=0])
```

```
data$STORIES[data$STORIES>=14] <- floor(median(data$STORIES[data$STORIES<14]))
```

The height of buildings in Washington is limited by the Height of Buildings Act. Tallest residential building in Washington, D.C. Tallest building completed in the city in the 2000s has 14 floors

```
#write.csv(data,'../data/pre_data.csv')
```

##Encode

```
data$AC <- factor(data$AC,level=c('Y','N'), label=c(1,0))




data$NATIONALGRID <- as.numeric(data$NATIONALGRID)
data$ASSESSMENT_NBHD <- as.factor(data$ASSESSMENT_NBHD)
data$STYLE <- as.numeric(data$STYLE)
```

sample

4 preprocessing

sample

4.1 missing data

sample

4.2 outliers

sample

5 Smoothing methods

The main purpose of using the smoothing method is applying the spline and local regression rule into high dimensional data analyst. In this part, all the parameters automatically selected by s() (low rank thin plate(smoothing) spline), te() (tensor product smoothing spline) and ti()(interaction).

5.1 data preprocessing and modification

Smoothing method is a specific kind of linear(quadritic) method thus its data has more conditions than random-forest method and boosting method. Therefore it is necessary to preprocess the data for smoothing method first.

There are two kinds of data in the data set: numeric and categorical, and some variable can treat as numeric variable since it has significant priority between the levels.

Countinus numeric variables:
BATHROOM, ROOMS, REDRM, AYB, YR_RMDL, EYB, STORIES, STYLE, GBA, SALEDATE, FIREPLACES, LANDAREA, ZIPCODE, LATITUDE, LONGTITUDE, GENUSU_TRACT

All the variables above are obviouse countinus numeric variables without missing or NA data. Consider the large data size, make very variables into smoothing spline it help will increase the prediction accurace.

numeric variables tranded from string: GRADE, CNDTN

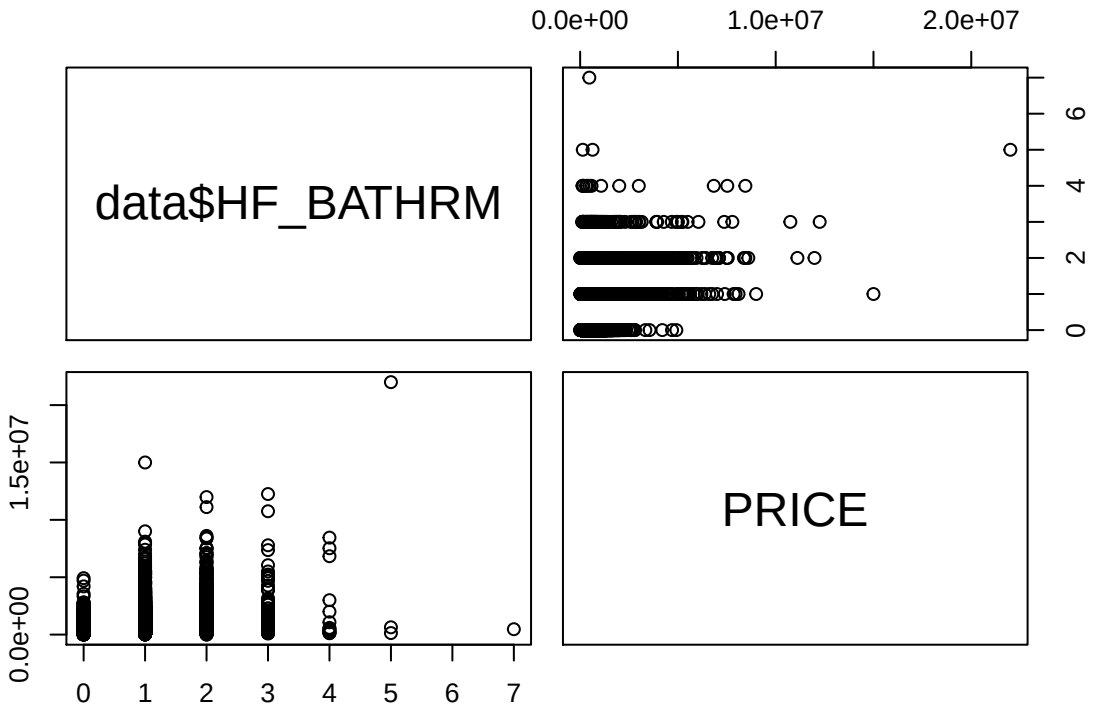
These two variables were saved as string in original dataset, but they actually present the quality of house, which showed have proorty for different factors. Thus we transfer these variables to numeric.

```
data$GRADE <- as.numeric(factor(data$GRADE,level=c('Low Quality', 'Fair Quality', 'Average', 'Above Ave
```

```
data$CNDTN <- as.numeric(factor(data$CNDTN,level=c('Poor', 'Fair', 'Average', 'Good', 'Very Good','Excel
```

HF_BATHROOM: Since the data levels of half bath room is only 8, and basing on the pairs plot.

```
pairs(~data$HF_BATHRM + PRICE, data = data)
```



Only 4 of the levels actually have most of the data, thus at first try to treat this variable as categorithon.

HOWever, a error will be reported as 'Error in predict.gam(gam.object, test) : 7 not in original fit', this error is frequently occur when we predict categorical variables.

NOT_IN_ORIGINAL_FIT

A frequently occur error when predict categorical variables. The main reason occur this error is categorical factor may not obvisu in every fold. Thus case if the fatcor not exist in 'train' fold but exist in test fold, the trained smooth model won't have a estimate parameter for that factor. This is the reason case the r carsh. The way we deal with this kind of problem is replacethe 'rare show up factor' thet not show up in every fold to some comom factors.

Howevery, for HF_BATHRM variable, the pair graph clearly shows that there should be a quadratic relationship between HF_BATHRM and PRICE, so treat HF_BATHRM as a numeric variable who has quadratic relationship.

Categorical variables:
AC, STRUCT, KITCKENS, USECODE, GENsus_TRACT,WARD,QUADRANT,ASSESSMENT_NBHD

All the Categorical variables above do not have missing data or NA, and all of their factors exist in every fold.

Heat: it is a obvious categoricall varible, but NOT_IN_ORIGINAL_FIT error exsit, do following transfer to avoid it.

```
data$HEAT <- as.character(data$HEAT)
data$HEAT[data$HEAT=='Air-oil'|
                data$HEAT=='Electric Rad'|
                 data$HEAT=='Evp Cool'|
                 data$HEAT=='Gravity Furnac'|
                 data$HEAT=='Ind Unit'|
                 data$HEAT=='No Data'|
                 data$HEAT=='Wall Furnace'
                 ] <- sample(data$HEAT[data$HEAT!='Air-oil'& data$HEAT=='Electric Rad'
                                       & data$HEAT!='Evp Cool'&
                                         data$HEAT!='Gravity Furnac'&
                                         data$HEAT!='Ind Unit'&
                                         data$HEAT!='No Data'&
                                         data$HEAT!='Wall Furnace'
                                       ],size = 8)
```

```
## Warning in data$HEAT[data$HEAT == "Air-oil" | data$HEAT == "Electric Rad"
## | : number of items to replace is not a multiple of replacement length
```

```
data$HEAT<-as.factor(data$HEAT)
```

Where we replace the rare obvisou data as smple from other data, random drawn exsit here, maycase every estimate lead to slight different k-variances!

And use the similar idea to preprocessing EXTWALL/ROOF/INTWALL

```
#EXTWALL
data$EXTWALL[data$EXTWALL == 'Adobe'| data$EXTWALL == 'Default'| data$EXTWALL == 'Plywoo
data$EXTWALL<-as.factor(data$EXTWALL)



data$QUADRANT[data$QUADRANT == ''] <- sample(data$QUADRANT[data$QUADRANT != ''],size = 6

data$INTWALL[data$INTWALL == 'Vinyl Comp'] <- 'Carpet'
```

Since there are too many missing data, we avoid estimate ASSESSMENT_SUBNBHD and CENUSU_BLOOK.
FULLADDRESS & NATIONALGRID has too many ovsivations that can not treat as categorical, but they are also meanless as numerical, so drop them out of estimate.

5.2 estimate single variable

Firstly build a linear regrassion model for all of the numeric variables as smooth spline. If the variable is importanmt in linear model(variable has less p-value), also means it will be important in predition model.

```
t1 <- gam(PRICE~ s(BATHRM) + s(ROOMS) + s(BEDRM)+ s(AYB)  + s(YR_RMDL)+ s(EYB) + s(STORI
          (STYLE) + s(FIREPLACES) + s(LANDAREA) + s(ZIPCODE) + s(LATITUDE) + s(LONGITUDE)
summary(t1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## PRICE ~ s(BATHRM) + s(ROOMS) + s(BEDRM) + s(AYB) + s(YR_RMDL) +
##     s(EYB) + s(STORIES) + s(STYLE) + s(FIREPLACES) + s(LANDAREA) +
##     s(ZIPCODE) + s(LATITUDE) + s(LONGITUDE) + s(CENSUS_TRACT) +
##     GRADE + CNDTN
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -83106      14916  -5.572 2.54e-08 ***
## GRADE          68076       2276  29.910  < 2e-16 ***
## CNDTN         106001       2800  37.860  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                   edf Ref.df       F  p-value
```

```
## s(BATHRM)       8.896  8.993 448.448  < 2e-16 ***
## s(ROOMS)        9.000  9.000  91.706  < 2e-16 ***
## s(BEDRM)        8.835  8.982  45.181  < 2e-16 ***
## s(AYB)          8.965  8.999 284.705  < 2e-16 ***
## s(YR_RMDL)      7.188  8.025  70.933  < 2e-16 ***
## s(EYB)          8.691  8.951 354.451  < 2e-16 ***
## s(STORIES)      1.168  1.314   2.829 0.077502 .
## s(STYLE)        5.867  6.697   3.990 0.000332 ***
## s(FIREPLACES)   8.957  8.999 153.251  < 2e-16 ***
## s(LANDAREA)     8.981  9.000 525.509  < 2e-16 ***
## s(ZIPCODE)      8.896  8.994  20.553  < 2e-16 ***
## s(LATITUDE)     8.875  8.995  59.982  < 2e-16 ***
## s(LONGITUDE)    8.714  8.977  41.996  < 2e-16 ***
## s(CENSUS_TRACT) 8.705  8.971  34.638  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.757   Deviance explained = 75.8%
## GCV = 8.1506e+10  Scale est. = 8.127e+10  n = 39520
```

By the p-value of summary, Stories and style has significant larger p-value than others, so propobaly we have to drop these two variables from model.

Do the same for categorical variables.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## PRICE ~ s(BATHRM) + s(ROOMS) + s(BEDRM) + HEAT + EXTWALL + ROOF +
##     INTWALL + AC + STRUCT + KITCHENS + USECODE + ASSESSMENT_NBHD +
##     WARD + QUADRANT
##
## Parametric coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                              672268.8   300734.0   2.235
## HEATAir Exchng                           -40148.4   132934.4  -0.302
## HEATElec Base Brd                         95311.5    89119.3   1.069
## HEATElectric Rad                          17790.5    76372.0   0.233
## HEATForced Air                            51312.3    69408.0   0.739
## HEATHot Water Rad                         28430.0    69363.6   0.410
## HEATHt Pump                               54358.4    71001.4   0.766
## HEATWarm Cool                             14253.4    69466.4   0.205
## HEATWater Base Brd                       103026.8    82234.6   1.253
## EXTWALLBrick Veneer                       49864.1    25798.9   1.933
## EXTWALLBrick/Siding                       -8294.0    20642.2  -0.402
```

```
## EXTWALLBrick/Stone                58316.1    29029.3   2.009
## EXTWALLBrick/Stucco                8959.9    27796.1   0.322
## EXTWALLCommon Brick                 906.4    19725.9   0.046
## EXTWALLConcrete                   48238.5    72181.1   0.668
## EXTWALLConcrete Block            -77009.6    78454.8  -0.982
## EXTWALLFace Brick                 17586.0    30042.4   0.585
## EXTWALLHardboard                 151971.1    52309.1   2.905
## EXTWALLMetal Siding               43056.9    80267.0   0.536
## EXTWALLShingle                    -2125.8    26339.3  -0.081
## EXTWALLStone                      67354.5    28933.1   2.328
## EXTWALLStone Veneer               24578.4    38338.9   0.641
## EXTWALLStone/Siding               32231.0    32591.9   0.989
## EXTWALLStone/Stucco              131446.9    36565.5   3.595
## EXTWALLStucco                     49347.3    21800.0   2.264
## EXTWALLStucco Block              -47356.2    84744.0  -0.559
## EXTWALLVinyl Siding              -13598.2    20682.4  -0.657
## EXTWALLWood Siding                20234.9    21176.7   0.956
## ROOFClay Tile                     22325.5    23254.7   0.960
## ROOFComp Shingle                 -24810.8     6365.9  -3.897
## ROOFComposition Ro                27315.9    50656.9   0.539
## ROOFConcrete                     -93601.0   339209.5  -0.276
## ROOFConcrete Tile               -416575.3   195117.2  -2.135
## ROOFMetal- Cpr                    -1136.1    94591.1  -0.012
## ROOFMetal- Pre                    30907.6    37874.2   0.816
## ROOFMetal- Sms                   -17463.3     5511.1  -3.169
## ROOFNeopren                       93724.3    14936.8   6.275
## ROOFShake                        -15321.5    20560.1  -0.745
## ROOFShingle                      -41279.9    25529.4  -1.617
## ROOFSlate                         33353.8     8333.4   4.002
## ROOFTypical                      -50867.2    43097.0  -1.180
## ROOFWater Proof                 -140697.9   239877.0  -0.587
## ROOFWood- FS                    -181961.4   196199.8  -0.927
## INTWALLCeramic Tile              -45635.7    66066.7  -0.691
## INTWALLDefault                    52281.9    70136.9   0.745
## INTWALLHardwood                   42048.5    10357.9   4.060
## INTWALLHardwood/Carp               6136.9    11059.9   0.555
## INTWALLLt Concrete               -36165.6    58486.8  -0.618
## INTWALLParquet                    46427.8   138816.3   0.334
## INTWALLResiliant                  94870.5   170277.7   0.557
## INTWALLTerrazo                  1484592.1   339736.3   4.370
## INTWALLVinyl Sheet                 5833.1   196121.9   0.030
## INTWALLWood Floor                 11525.4    12428.5   0.927
## ACO                             -105549.3     5526.4 -19.099
## STRUCTMulti                     -414473.7   246083.6  -1.684
## STRUCTRow End                   -300103.2   240323.3  -1.249
## STRUCTRow Inside                -313382.6   240320.8  -1.304
```

```
## STRUCTSemi-Detached                          -276479.8   240320.0  -1.150
## STRUCTSingle                                 -227011.1   239951.8  -0.946
## STRUCTTown End                               -285877.6   244101.3  -1.171
## STRUCTTown Inside                            -303725.9   242102.3  -1.255
## KITCHENS                                      99370.4     8393.1  11.840
## USECODE                                       -5557.8    13566.3  -0.410
## ASSESSMENT_NBHDAmerican University           142540.8    22762.8   6.262
## ASSESSMENT_NBHDAnacostia                      73008.4   158330.8   0.461
## ASSESSMENT_NBHDBarry Farms                    71288.2   162216.7   0.439
## ASSESSMENT_NBHDBerkley                       364722.8    27066.5  13.475
## ASSESSMENT_NBHDBrentwood                     -22490.7    38771.1  -0.580
## ASSESSMENT_NBHDBrightwood                    -48221.0    16177.9  -2.981
## ASSESSMENT_NBHDBrookland                      85772.1    31251.0   2.745
## ASSESSMENT_NBHDBurleith                       92554.1    35243.7   2.626
## ASSESSMENT_NBHDCapitol Hill                  393238.4    39214.9  10.028
## ASSESSMENT_NBHDCentral-tri 1                 523326.7    65420.1   7.999
## ASSESSMENT_NBHDChevy Chase                   145797.7    16745.3   8.707
## ASSESSMENT_NBHDChillum                       -74130.8    23916.0  -3.100
## ASSESSMENT_NBHDCleveland Park                474769.7    25333.6  18.741
## ASSESSMENT_NBHDColonial Village              -37926.0    26871.7  -1.411
## ASSESSMENT_NBHDColumbia Heights              70870.5     21261.3   3.333
## ASSESSMENT_NBHDCongress Heights              15111.8    159336.2   0.095
## ASSESSMENT_NBHDCrestwood                     102210.2    23985.3   4.261
## ASSESSMENT_NBHDDeanwood                       39861.4    43437.7   0.918
## ASSESSMENT_NBHDEckington                     116859.8    30526.0   3.828
## ASSESSMENT_NBHDFoggy Bottom                   -5430.4    45974.8  -0.118
## ASSESSMENT_NBHDForest Hills                  213785.9    27559.4   7.757
## ASSESSMENT_NBHDFort Dupont Park               39573.1    45442.4   0.871
## ASSESSMENT_NBHDFort Lincoln                  137781.2    36044.0   3.823
## ASSESSMENT_NBHDFoxhall                       146229.2    31655.3   4.619
## ASSESSMENT_NBHDGarfield                      346608.1    31085.2  11.150
## ASSESSMENT_NBHDGeorgetown                    581587.6    31428.0  18.505
## ASSESSMENT_NBHDGlover Park                   224561.3    26388.9   8.510
## ASSESSMENT_NBHDHawthorne                      21673.3    33858.4   0.640
## ASSESSMENT_NBHDHillcrest                      49392.7    45936.8   1.075
## ASSESSMENT_NBHDKalorama                      721820.2    34522.0  20.909
## ASSESSMENT_NBHDKent                          294046.1    26266.8  11.195
## ASSESSMENT_NBHDLedroit Park                  105699.3    26354.8   4.011
## ASSESSMENT_NBHDLily Ponds                    108649.9    44960.8   2.417
## ASSESSMENT_NBHDMarshall Heights               47199.3    46408.8   1.017
## ASSESSMENT_NBHDMassachusetts Avenue Heights  891848.0    42317.6  21.075
## ASSESSMENT_NBHDMichigan Park                 -39780.0    35406.4  -1.124
## ASSESSMENT_NBHDMt. Pleasant                  297436.3    26812.2  11.093
## ASSESSMENT_NBHDNorth Cleveland Park          186530.4    26441.7   7.054
## ASSESSMENT_NBHDObservatory Circle            369950.5    31562.6  11.721
## ASSESSMENT_NBHDOld City 1                    238648.5    37497.9   6.364
```

```
## ASSESSMENT_NBHDOld City 2                204689.0    24402.9   8.388
## ASSESSMENT_NBHDPalisades                 243664.2    25689.6   9.485
## ASSESSMENT_NBHDPetworth                   15380.7    15394.6   0.999
## ASSESSMENT_NBHDRandle Heights             76445.0   159410.3   0.480
## ASSESSMENT_NBHDRiggs Park                -20490.4    29015.0  -0.706
## ASSESSMENT_NBHDShepherd Heights          -44171.9    22178.7  -1.992
## ASSESSMENT_NBHDSouthwest Waterfront      177719.8    51703.5   3.437
## ASSESSMENT_NBHDSpring Valley             294692.9    26115.8  11.284
## ASSESSMENT_NBHDTakoma Park                -5263.1    25522.4  -0.206
## ASSESSMENT_NBHDTrinidad                   55296.9    32751.0   1.688
## ASSESSMENT_NBHDWakefield                 127897.8    32358.9   3.952
## ASSESSMENT_NBHDWesley Heights            355123.6    27997.0  12.684
## ASSESSMENT_NBHDWoodley                   345489.8    39407.5   8.767
## ASSESSMENT_NBHDWoodridge                 -57116.2    32290.2  -1.769
## WARDWard 2                               328606.5    21927.4  14.986
## WARDWard 3                                52828.5    22614.6   2.336
## WARDWard 4                                -7589.8    17471.0  -0.434
## WARDWard 5                               -23997.2    17314.7  -1.386
## WARDWard 6                               -30658.8    23349.7  -1.313
## WARDWard 7                              -185084.3    31753.2  -5.829
## WARDWard 8                              -179587.2   155963.0  -1.151
## QUADRANTNW                                47129.8    20616.3   2.286
## QUADRANTSE                                18243.1     8848.5   2.062
## QUADRANTSW                                -9318.7    23558.1  -0.396
##                                         Pr(>|t|)
## (Intercept)                             0.025395 *
## HEATAir Exchng                          0.762641
## HEATElec Base Brd                       0.284859
## HEATElectric Rad                        0.815805
## HEATForced Air                          0.459738
## HEATHot Water Rad                       0.681904
## HEATHt Pump                             0.443922
## HEATWarm Cool                           0.837429
## HEATWater Base Brd                      0.210271
## EXTWALLBrick Veneer                     0.053268 .
## EXTWALLBrick/Siding                     0.687833
## EXTWALLBrick/Stone                      0.044558 *
## EXTWALLBrick/Stucco                     0.747195
## EXTWALLCommon Brick                     0.963351
## EXTWALLConcrete                         0.503947
## EXTWALLConcrete Block                   0.326313
## EXTWALLFace Brick                       0.558300
## EXTWALLHardboard                        0.003672 **
## EXTWALLMetal Siding                     0.591671
## EXTWALLShingle                          0.935673
## EXTWALLStone                            0.019920 *
```

```
## EXTWALLStone Veneer                          0.521473
## EXTWALLStone/Siding                          0.322704
## EXTWALLStone/Stucco                          0.000325 ***
## EXTWALLStucco                                0.023602 *
## EXTWALLStucco Block                          0.576291
## EXTWALLVinyl Siding                          0.510878
## EXTWALLWood Siding                           0.339316
## ROOFClay Tile                                0.337038
## ROOFComp Shingle                             9.74e-05 ***
## ROOFComposition Ro                           0.589729
## ROOFConcrete                                 0.782597
## ROOFConcrete Tile                            0.032767 *
## ROOFMetal- Cpr                               0.990418
## ROOFMetal- Pre                               0.414471
## ROOFMetal- Sms                               0.001532 **
## ROOFNeopren                                  3.54e-10 ***
## ROOFShake                                    0.456152
## ROOFShingle                                  0.105895
## ROOFSlate                                    6.28e-05 ***
## ROOFTypical                                  0.237890
## ROOFWater Proof                              0.557515
## ROOFWood- FS                                 0.353710
## INTWALLCeramic Tile                          0.489725
## INTWALLDefault                               0.456018
## INTWALLHardwood                              4.93e-05 ***
## INTWALLHardwood/Carp                         0.578982
## INTWALLLt Concrete                           0.536345
## INTWALLParquet                               0.738038
## INTWALLResiliant                             0.577427
## INTWALLTerrazo                               1.25e-05 ***
## INTWALLVinyl Sheet                           0.976273
## INTWALLWood Floor                            0.353757
## AC0                                           < 2e-16 ***
## STRUCTMulti                                  0.092136 .
## STRUCTRow End                                0.211765
## STRUCTRow Inside                             0.192235
## STRUCTSemi-Detached                          0.249959
## STRUCTSingle                                 0.344119
## STRUCTTown End                               0.241548
## STRUCTTown Inside                            0.209655
## KITCHENS                                      < 2e-16 ***
## USECODE                                      0.682047
## ASSESSMENT_NBHDAmerican University           3.84e-10 ***
## ASSESSMENT_NBHDAnacostia                     0.644720
## ASSESSMENT_NBHDBarry Farms                   0.660329
## ASSESSMENT_NBHDBerkley                        < 2e-16 ***
```

```
## ASSESSMENT_NBHDBrentwood                     0.561858
## ASSESSMENT_NBHDBrightwood                    0.002878 **
## ASSESSMENT_NBHDBrookland                     0.006061 **
## ASSESSMENT_NBHDBurleith                      0.008640 **
## ASSESSMENT_NBHDCapitol Hill                   < 2e-16 ***
## ASSESSMENT_NBHDCentral-tri 1                 1.28e-15 ***
## ASSESSMENT_NBHDChevy Chase                    < 2e-16 ***
## ASSESSMENT_NBHDChillum                       0.001939 **
## ASSESSMENT_NBHDCleveland Park                 < 2e-16 ***
## ASSESSMENT_NBHDColonial Village              0.158142
## ASSESSMENT_NBHDColumbia Heights              0.000859 ***
## ASSESSMENT_NBHDCongress Heights              0.924441
## ASSESSMENT_NBHDCrestwood                     2.04e-05 ***
## ASSESSMENT_NBHDDeanwood                      0.358798
## ASSESSMENT_NBHDEckington                     0.000129 ***
## ASSESSMENT_NBHDFoggy Bottom                  0.905976
## ASSESSMENT_NBHDForest Hills                  8.89e-15 ***
## ASSESSMENT_NBHDFort Dupont Park              0.383847
## ASSESSMENT_NBHDFort Lincoln                  0.000132 ***
## ASSESSMENT_NBHDFoxhall                       3.86e-06 ***
## ASSESSMENT_NBHDGarfield                       < 2e-16 ***
## ASSESSMENT_NBHDGeorgetown                     < 2e-16 ***
## ASSESSMENT_NBHDGlover Park                    < 2e-16 ***
## ASSESSMENT_NBHDHawthorne                     0.522101
## ASSESSMENT_NBHDHillcrest                     0.282278
## ASSESSMENT_NBHDKalorama                       < 2e-16 ***
## ASSESSMENT_NBHDKent                           < 2e-16 ***
## ASSESSMENT_NBHDLedroit Park                  6.07e-05 ***
## ASSESSMENT_NBHDLily Ponds                    0.015673 *
## ASSESSMENT_NBHDMarshall Heights              0.309144
## ASSESSMENT_NBHDMassachusetts Avenue Heights   < 2e-16 ***
## ASSESSMENT_NBHDMichigan Park                 0.261221
## ASSESSMENT_NBHDMt. Pleasant                   < 2e-16 ***
## ASSESSMENT_NBHDNorth Cleveland Park          1.76e-12 ***
## ASSESSMENT_NBHDObservatory Circle             < 2e-16 ***
## ASSESSMENT_NBHDOld City 1                    1.98e-10 ***
## ASSESSMENT_NBHDOld City 2                     < 2e-16 ***
## ASSESSMENT_NBHDPalisades                      < 2e-16 ***
## ASSESSMENT_NBHDPetworth                      0.317754
## ASSESSMENT_NBHDRandle Heights                0.631551
## ASSESSMENT_NBHDRiggs Park                    0.480069
## ASSESSMENT_NBHDShepherd Heights              0.046418 *
## ASSESSMENT_NBHDSouthwest Waterfront          0.000588 ***
## ASSESSMENT_NBHDSpring Valley                  < 2e-16 ***
## ASSESSMENT_NBHDTakoma Park                   0.836623
## ASSESSMENT_NBHDTrinidad                      0.091341 .
```

```
## ASSESSMENT_NBHDWakefield                    7.75e-05 ***
## ASSESSMENT_NBHDWesley Heights                < 2e-16 ***
## ASSESSMENT_NBHDWoodley                       < 2e-16 ***
## ASSESSMENT_NBHDWoodridge                    0.076928 .
## WARDWard 2                                   < 2e-16 ***
## WARDWard 3                                  0.019494 *
## WARDWard 4                                  0.663986
## WARDWard 5                                  0.165772
## WARDWard 6                                  0.189182
## WARDWard 7                                  5.62e-09 ***
## WARDWard 8                                  0.249545
## QUADRANTNW                                  0.022257 *
## QUADRANTSE                                  0.039241 *
## QUADRANTSW                                  0.692431
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df      F p-value
## s(BATHRM) 8.900  8.993 660.82  <2e-16 ***
## s(ROOMS)  8.993  9.000 159.77  <2e-16 ***
## s(BEDRM)  8.916  8.995  68.06  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.657   Deviance explained = 65.9%
## GCV = 1.1529e+11  Scale est. = 1.1484e+11  n = 39520
```

By the defination of categorical estimate, r treat every factors as an independent variable, but in our prediction model later we can not only a part of the categorical variable. Since for most of categorical variables their factors' p-values are pretty different from each orther. We can not decide witch variables we want.

6 Random Forests

sample

7 Boosting

sample

8 Aditional methods

sample

9 Statistical Conclusions

sample

10 Future work

sample

11 Contribution

sample

12 Appendix

sample