Saratoga Houses Price Modeling Strategies for the Local Taxing Authority

Bao Doquang, Dhwanit Agarwal, Akksay Singh and Shristi Singh

March 13, 2020

Classwork:

```
library(tidyverse, quietly = TRUE)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.2

## -- Attaching packages -----------------------------------------------------------------------

## <U+2713> ggplot2 3.2.1     <U+2713> purrr   0.3.3
## <U+2713> tibble  2.1.3     <U+2713> dplyr   0.8.4
## <U+2713> tidyr   1.0.0     <U+2713> stringr 1.4.0
## <U+2713> readr   1.3.1     <U+2713> forcats 0.4.0

## Warning: package 'dplyr' was built under R version 3.6.3

## -- Conflicts --------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr, quietly = TRUE)
library(mosaic, quietly = TRUE)
```

```
## Warning: package 'mosaic' was built under R version 3.6.2

## Warning: package 'ggstance' was built under R version 3.6.2

##
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh

##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")

## Warning: package 'mosaicData' was built under R version 3.6.2

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
##
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##
##     mean

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following object is masked from 'package:purrr':
##
##     cross

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
library(FNN, quietly = TRUE)
```

```
## Warning: package 'FNN' was built under R version 3.6.2
```

```
library(foreach, quietly = TRUE)
```

```
## Warning: package 'foreach' was built under R version 3.6.3

##
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

```
data(SaratogaHouses)

summary(SaratogaHouses)
```

```
##      price           lotSize              age           landValue
##  Min.   :  5000   Min.   : 0.0000   Min.   :  0.00   Min.   :   200
##  1st Qu.:145000   1st Qu.: 0.1700   1st Qu.: 13.00   1st Qu.: 15100
##  Median :189900   Median : 0.3700   Median : 19.00   Median : 25000
##  Mean   :211967   Mean   : 0.5002   Mean   : 27.92   Mean   : 34557
##  3rd Qu.:259000   3rd Qu.: 0.5400   3rd Qu.: 34.00   3rd Qu.: 40200
##  Max.   :775000   Max.   :12.2000   Max.   :225.00   Max.   :412600
##    livingArea     pctCollege       bedrooms       fireplaces       bathrooms
##  Min.   : 616   Min.   :20.00   Min.   :1.000   Min.   :0.0000   Min.   :0.0
##  1st Qu.:1300   1st Qu.:52.00   1st Qu.:3.000   1st Qu.:0.0000   1st Qu.:1.5
##  Median :1634   Median :57.00   Median :3.000   Median :1.0000   Median :2.0
##  Mean   :1755   Mean   :55.57   Mean   :3.155   Mean   :0.6019   Mean   :1.9
##  3rd Qu.:2138   3rd Qu.:64.00   3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:2.5
##  Max.   :5228   Max.   :82.00   Max.   :7.000   Max.   :4.0000   Max.   :4.5
##      rooms                   heating           fuel
##  Min.   : 2.000   hot air         :1121   gas     :1197
##  1st Qu.: 5.000   hot water/steam: 302   electric: 315
##  Median : 7.000   electric        : 305   oil     : 216
##  Mean   : 7.042
##  3rd Qu.: 8.250
##  Max.   :12.000
##                sewer      waterfront newConstruction centralAir
##  septic           : 503   Yes:  15   Yes:  81        Yes: 635
##  public/commercial:1213   No :1713   No :1647        No :1093
##  none             :  12
##
##
##
```

```
#Defining models

# Baseline model
lm_small = lm(price ~ bedrooms + bathrooms + lotSize, data=SaratogaHouses)

# 11 main effects
lm_medium = lm(price ~ lotSize + age + livingArea + pctCollege + bedrooms +
                 fireplaces + bathrooms + rooms + heating + fuel + centralAir, data=SaratogaHouses)

# Sometimes it's easier to name the variables we want to leave out
# The command below yields exactly the same model.
# the dot (.) means "all variables not named"
# the minus (-) means "exclude this variable"
lm_medium2 = lm(price ~ . - sewer - waterfront - landValue - newConstruction, data=SaratogaHouses)

coef(lm_medium)
```

```
##           (Intercept)               lotSize                   age
##           28627.73165            9350.45188              47.54722
##            livingArea            pctCollege              bedrooms
##              91.86974             296.50809          -15630.71950
##            fireplaces             bathrooms                 rooms
##             985.06117           22006.97108            3259.11923
## heatinghot water/steam       heatingelectric          fuelelectric
##           -9429.79463           -3609.98574          -12094.12195
##               fueloil           centralAirNo
##           -8873.13971          -17112.81908
```

```
coef(lm_medium2)
```

```
##           (Intercept)               lotSize                   age
##            28627.73165            9350.45188             47.54722
##             livingArea            pctCollege             bedrooms
##               91.86974             296.50809         -15630.71950
##             fireplaces             bathrooms                rooms
##              985.06117           22006.97108           3259.11923
## heatinghot water/steam        heatingelectric          fuelelectric
##            -9429.79463           -3609.98574         -12094.12195
##                fueloil            centralAirNo
##            -8873.13971          -17112.81908
```

```
# All interactions
# the ()^2 says "include all pairwise interactions"
lm_big = lm(price ~ (. - sewer - waterfront - landValue - newConstruction)^2, data=SaratogaHouses)


####
# Compare out-of-sample predictive performance
####

# Split into training and testing sets
n = nrow(SaratogaHouses) # number of rows
n_train = round(0.8*n)  # round to nearest integer
n_test = n - n_train
train_cases = sample.int(n, n_train, replace=FALSE)
test_cases = setdiff(1:n, train_cases)
saratoga_train = SaratogaHouses[train_cases,]
saratoga_test = SaratogaHouses[test_cases,]

# Fit to the training data
lm1 = lm(price ~ lotSize + bedrooms + bathrooms, data=saratoga_train)
lm2 = lm(price ~ . - sewer - waterfront - landValue - newConstruction, data=saratoga_train)
lm3 = lm(price ~ (. - sewer - waterfront - landValue - newConstruction)^2, data=saratoga_train)

# Predictions out of sample
yhat_test1 = predict(lm1, saratoga_test)
yhat_test2 = predict(lm2, saratoga_test)
yhat_test3 = predict(lm3, saratoga_test)
```

```
## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading
```

```
rmse = function(y, yhat) {
  sqrt( mean( (y - yhat)^2 ) )
}

# Root mean-squared prediction error
rmse(saratoga_test$price, yhat_test1)
```

```
## [1] 84041.02
```

```
rmse(saratoga_test$price, yhat_test2)
```

```
## [1] 69905.94
```

```
rmse(saratoga_test$price, yhat_test3)
```

```
## [1] 73548.62
```

```r
# easy averaging over train/test splits

n_train = round(0.8*n)  # round to nearest integer
n_test = n - n_train

rmse_vals = do(100)*{

  # re-split into train and test cases with the same sample sizes
  train_cases = sample.int(n, n_train, replace=FALSE)
  test_cases = setdiff(1:n, train_cases)
  saratoga_train = SaratogaHouses[train_cases,]
  saratoga_test = SaratogaHouses[test_cases,]

  # Fit to the training data
  lm1 = lm(price ~ lotSize + bedrooms + bathrooms, data=saratoga_train)
  lm2 = lm(price ~ . - sewer - waterfront - landValue - newConstruction, data=saratoga_train)
  lm3 = lm(price ~ (. - sewer - waterfront - landValue - newConstruction)^2, data=saratoga_train)

  lm_dominate = lm(price ~ lotSize + age + livingArea + pctCollege +
                     bedrooms + fireplaces + bathrooms + rooms + heating + fuel +
                     centralAir + lotSize:heating + livingArea:rooms + newConstruction + livingArea:new

  # Predictions out of sample
  yhat_test1 = predict(lm1, saratoga_test)
  yhat_test2 = predict(lm2, saratoga_test)
  yhat_test3 = predict(lm3, saratoga_test)
  yhat_test4 = predict(lm_dominate, saratoga_test)

  c(rmse(saratoga_test$price, yhat_test1),
    rmse(saratoga_test$price, yhat_test2),
    rmse(saratoga_test$price, yhat_test3),
    rmse(saratoga_test$price, yhat_test4))
}
```

```
## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading
## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading
## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading

## Warning in predict.lm(lm3, saratoga_test): prediction from a rank-deficient fit
## may be misleading
```

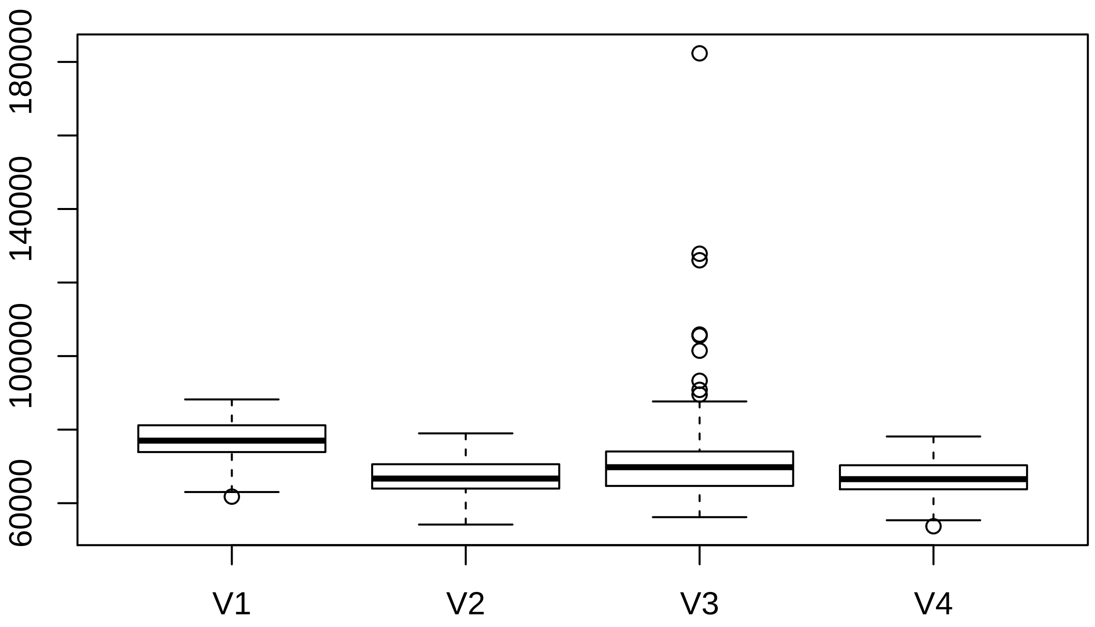
```
rmse_vals
```

```
##           V1       V2        V3       V4
## 1   73523.48 61536.94  71437.06 61469.28
## 2   72835.41 64946.46  66828.71 64120.18
## 3   71310.51 63708.26  61220.46 63146.32
## 4   77102.25 67346.51  68824.50 66534.45
## 5   77452.59 67088.88 182330.72 67542.95
## 6   79090.54 66096.99  69180.13 65857.37
## 7   79567.02 67383.58  71491.48 66626.54
## 8   77376.38 64827.83  63550.53 64537.97
## 9   82371.20 74063.84  75281.09 74301.34
## 10  76353.27 64234.64  66196.51 64998.41
## 11  76468.30 66568.55  66862.14 65637.30
## 12  76663.25 66692.13  66728.55 66524.43
## 13  78385.35 69726.74  69850.60 70250.24
## 14  73036.10 64009.29  63392.61 63541.01
## 15  83712.93 71347.78  77579.28 71086.50
## 16  70773.10 61251.29  60803.87 60761.11
## 17  88214.56 74612.46  84059.13 73794.17
## 18  75003.35 66081.08  66550.04 66384.72
## 19  77361.51 67385.88  76454.07 67171.12
## 20  75364.86 68109.36  70309.52 70399.85
## 21  71977.81 55379.23  58221.56 55395.69
## 22  79264.99 65154.87  65121.01 64528.71
## 23  83833.52 72194.41  71681.70 72653.86
## 24  73001.69 63554.07  65506.29 63462.50
## 25  82791.05 72047.71  69715.21 71463.47
## 26  75664.39 66417.93  67572.45 66947.67
## 27  74336.74 60831.28  90809.49 59914.08
## 28  84317.25 72171.02 127836.22 71306.17
## 29  77759.16 66995.24  70667.76 68498.66
## 30  83499.13 73518.60  71185.62 73139.94
## 31  72408.19 61316.38  63528.13 62025.99
## 32  85795.50 72180.80  74680.14 73664.28
## 33  63979.51 58510.69  63004.33 59456.83
## 34  81572.27 72312.51  93281.59 71554.34
## 35  73421.38 65113.82  64262.70 65215.72
## 36  84321.09 70922.63  72478.74 70050.08
## 37  82348.99 71765.48  89520.82 71034.58
## 38  76973.41 64142.27  68709.90 64571.19
## 39  87275.47 72789.30 126059.17 71973.05
## 40  81294.01 70224.24  73859.22 69547.35
## 41  73119.09 63960.91  62569.86 64224.87
## 42  72759.72 61305.63  61479.98 60709.03
## 43  80549.61 68296.86  69988.49 67529.26
## 44  78357.65 66391.88  69658.97 66136.36
## 45  77614.38 65930.41  64130.56 65031.00
## 46  82659.44 71469.54  74267.58 71211.41
## 47  74726.04 66697.82  69906.95 66216.14
## 48  81153.30 72927.22 101472.97 72683.21
## 49  78009.98 68244.55  68592.65 67818.37
## 50  75545.02 65215.61  67089.25 65274.89
## 51  82926.95 72972.80  71944.37 72346.12
## 52  76598.47 68229.85 105569.82 68493.18
## 53  64355.70 55832.81  56667.26 56428.04
## 54  73321.47 61641.91  62656.83 61208.45
## 55  70809.78 60231.91  58214.12 60223.01
## 56  63033.47 55859.52  62126.37 55964.03
## 57  70738.47 60854.43  63761.88 62908.66
## 58  75899.26 67296.15  75114.88 66927.98
## 59  83032.14 71576.84  72788.00 72562.04
## 60  75799.58 66813.93  68180.86 67305.18
## 61  75783.66 68183.84  66552.22 67814.33
## 62  82621.66 71514.76  75915.81 70872.29
## 63  61776.66 54181.75  56199.53 53728.20
## 64  71323.55 60546.26  60956.87 60736.77
## 65  66765.37 56101.38  61394.69 55359.65
## 66  72364.48 60547.06  59408.48 60059.84
## 67  82274.49 78981.64  78273.30 78138.92
## 68  75571.82 65204.44  65274.66 64529.42
## 69  78437.83 64912.29  87662.64 64385.95
## 70  74245.18 68728.45  68298.34 69167.01
## 71  76256.23 70405.12  69744.73 69663.33
## 72  81060.77 70785.66  72883.40 70384.46
## 73  74916.02 62136.76  62759.56 62545.72
## 74  75175.35 66102.05  70515.17 65663.23
## 75  76836.58 69787.06  76376.47 69723.14
## 76  76566.73 65846.92  68171.65 65012.31
## 77  78320.84 69270.04  71066.28 68553.23
## 78  70348.10 61531.58  60788.99 60795.95
## 79  81218.36 66571.87  72281.58 66230.06
## 80  82500.57 71481.78  72687.71 71444.71
## 81  71979.61 62418.00  76538.35 62513.90
## 82  80523.70 67833.60  73804.26 68916.42
## 83  80821.51 72328.90  73544.24 72974.21
## 84  77031.74 68302.75  73445.90 68066.56
## 85  86504.53 76261.38  73546.06 75490.54
## 86  78288.40 69601.12  67612.85 69156.38
## 87  74012.51 61807.81  62194.18 62165.26
## 88  75181.72 61063.60  69823.21 62728.54
## 89  81725.64 69368.93  73761.35 68918.83
## 90  78954.32 66912.05  66857.73 67035.55
## 91  74797.38 65307.05  76325.96 64552.46
## 92  77630.53 66194.00  69717.92 65423.82
## 93  86795.30 78062.65  75236.03 77093.23
## 94  73786.51 66499.63  65282.08 66063.17
## 95  83388.26 71893.51 105858.45 72203.78
## 96  78072.20 65063.38  63466.52 64603.63
## 97  80904.03 69679.68  73230.87 69371.58
## 98  81714.51 74957.95  77016.91 76530.61
## 99  67700.17 57613.33  77631.81 57608.32
## 100 74437.65 63983.78  61709.11 64021.64
```

```r
colMeans(rmse_vals)
```

```
##       V1       V2       V3       V4
## 77074.25 66742.85 72866.31 66625.38
```

```r
boxplot(rmse_vals)
```



Attempt at “hand-building” a model for price that outperforms the “medium” model that we considered in class by using combinations of transformations, polynomial terms, and interactions:

```r
str(SaratogaHouses)
```

```
## 'data.frame':    1728 obs. of  16 variables:
##  $ price          : int  132500 181115 109000 155000 86060 120000 153000 170000 90000 122900 ...
##  $ lotSize        : num  0.09 0.92 0.19 0.41 0.11 0.68 0.4 1.21 0.83 1.94 ...
##  $ age            : int  42 0 133 13 0 31 33 23 36 4 ...
##  $ landValue      : int  50000 22300 7300 18700 15000 14000 23300 14600 22200 21200 ...
##  $ livingArea     : int  906 1953 1944 1944 840 1152 2752 1662 1632 1416 ...
##  $ pctCollege     : int  35 51 51 51 51 22 51 35 51 44 ...
##  $ bedrooms       : int  2 3 4 3 2 4 4 4 3 3 ...
##  $ fireplaces     : int  1 0 1 1 0 1 1 1 0 0 ...
##  $ bathrooms      : num  1 2.5 1 1.5 1 1 1.5 1.5 1.5 1.5 ...
##  $ rooms          : int  5 6 8 5 3 8 8 9 8 6 ...
##  $ heating        : Factor w/ 3 levels "hot air","hot water/steam",..: 3 2 2 1 1 1 2 1 3 1 ...
##  $ fuel           : Factor w/ 3 levels "gas","electric",..: 2 1 1 1 1 1 3 3 2 1 ...
##  $ sewer          : Factor w/ 3 levels "septic","public/commercial",..: 1 1 2 1 2 1 1 1 1 3 ...
##  $ waterfront     : Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 2 2 2 ...
##  $ newConstruction: Factor w/ 2 levels "Yes","No": 2 2 2 2 1 2 2 2 2 2 ...
##  $ centralAir     : Factor w/ 2 levels "Yes","No": 2 2 2 2 1 2 2 2 2 2 ...
```

```r
# New variables for "hand-built" model

SaratogaHouses$ConstructionCost <- SaratogaHouses$price - SaratogaHouses$landValue
SaratogaHouses$waterfrontDummy <- ifelse(SaratogaHouses$waterfront == "yes", 1,0)
```

```r
SaratogaHouses$newConstructionDummy <- ifelse(SaratogaHouses$age == "yes", 1,0)
SaratogaHouses$centralAirDummy <- ifelse(SaratogaHouses$age == "yes", 1,0)

str(SaratogaHouses)
```

```
## 'data.frame':    1728 obs. of  20 variables:
##  $ price               : int  132500 181115 109000 155000 86060 120000 153000 170000 90000 122900 ..
##  $ lotSize             : num  0.09 0.92 0.19 0.41 0.11 0.68 0.4 1.21 0.83 1.94 ...
##  $ age                 : int  42 0 133 13 0 31 33 23 36 4 ...
##  $ landValue           : int  50000 22300 7300 18700 15000 14000 23300 14600 22200 21200 ...
##  $ livingArea          : int  906 1953 1944 1944 840 1152 2752 1662 1632 1416 ...
##  $ pctCollege          : int  35 51 51 51 51 22 51 35 51 44 ...
##  $ bedrooms            : int  2 3 4 3 2 4 4 4 3 3 ...
##  $ fireplaces          : int  1 0 1 1 0 1 1 1 0 0 ...
##  $ bathrooms           : num  1 2.5 1 1.5 1 1 1.5 1.5 1.5 1.5 ...
##  $ rooms               : int  5 6 8 5 3 8 8 9 8 6 ...
##  $ heating             : Factor w/ 3 levels "hot air","hot water/steam",..: 3 2 2 1 1 1 2 1 3 1 ...
##  $ fuel                : Factor w/ 3 levels "gas","electric",..: 2 1 1 1 1 1 3 3 2 1 ...
##  $ sewer               : Factor w/ 3 levels "septic","public/commercial",..: 1 1 2 1 2 1 1 1 1 3 ...
##  $ waterfront          : Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 2 2 2 ...
##  $ newConstruction     : Factor w/ 2 levels "Yes","No": 2 2 2 2 1 2 2 2 2 2 ...
##  $ centralAir          : Factor w/ 2 levels "Yes","No": 2 2 2 2 1 2 2 2 2 2 ...
##  $ ConstructionCost    : int  82500 158815 101700 136300 71060 106000 129700 155400 67800 101700 ...
##  $ waterfrontDummy     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ newConstructionDummy: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ centralAirDummy     : num  0 0 0 0 0 0 0 0 0 0 ...
```

```r
HeatingElectric <- SaratogaHouses[grep("electric", SaratogaHouses$heating), ]
#View(HeatingElectric)
#str(HeatingElectric)

HeatingSteam <- SaratogaHouses[grep("hot water/steam", SaratogaHouses$heating), ]
#View(HeatingSteam)
#str(HeatingSteam)

HeatingHotAir <- SaratogaHouses[grep("hot air", SaratogaHouses$heating), ]
#View(HeatingHotAir)
#str(HeatingHotAir)

FuelOil <- SaratogaHouses[grep("oil", SaratogaHouses$fuel), ]
#View(FuelOil)
#str(FuelOil)

FuelGas <- SaratogaHouses[grep("gas", SaratogaHouses$fuel), ]
#View(FuelGas)
#str(FuelGas)

FuelElectric <- SaratogaHouses[grep("electric", SaratogaHouses$fuel), ]
#View(FuelElectric)
#str(FuelElectric)

SewerSeptic <- SaratogaHouses[grep("septic", SaratogaHouses$sewer), ]
#View(SewerSeptic)
#str(SewerSeptic)
```

```r
SewerPublicCommercial <- SaratogaHouses[grep("public/commercial", SaratogaHouses$sewer), ]
#View(SewerPublicCommercial)
#str(SewerPublicCommercial)

SewerNone <- SaratogaHouses[grep("none", SaratogaHouses$sewer), ]
#View(SewerNone)
#str(SewerNone)

#Defining the models

#Baseline model

lm_medium = lm(price ~ lotSize + age + livingArea + pctCollege + bedrooms +
              fireplaces + bathrooms + rooms + heating + fuel + centralAir, data=SaratogaHouses)

#Hand-built Model

lm_handbuilt = lm(price ~ lotSize + age + livingArea + pctCollege + bedrooms +
                 fireplaces + bathrooms + rooms + heating + fuel + centralAir + ConstructionCost +
                 ConstructionCost*landValue + newConstructionDummy*landValue + newConstructionDummy*lotS
                 pctCollege*age + bathrooms*bedrooms, data = SaratogaHouses)


#Defining only the numerics of the train-test data sets
N = nrow(SaratogaHouses)
train = round(0.8*N)
test = (N-train)

#Defining the function
rmse = function(y, yhat) {
  sqrt( mean( (y - yhat)^2 ) )
}

#Rmse iterations
rmse1 <- NULL
rmse2 <- NULL


for (i in seq(1:00)){
  #Choosing data for training and testing
  train_cases = sample.int(N, train, replace=FALSE)
  test_cases = setdiff(1:N, train_cases)

  #Define the train-test data sets (for all X's and Y)
  saratoga_train = SaratogaHouses[train_cases,]
  saratoga_test = SaratogaHouses[test_cases,]

  #Training
  #Baseline model

  lm_medium = lm(price ~ lotSize + age + livingArea + pctCollege + bedrooms +
              fireplaces + bathrooms + rooms + heating + fuel + centralAir, data=saratoga_train)
  #Hand-built Model
  lm_handbuilt = lm(price ~ lotSize + age + livingArea + pctCollege + bedrooms +
                 fireplaces + bathrooms + rooms + heating + fuel + centralAir + ConstructionCost +
                 ConstructionCost*landValue + newConstructionDummy*landValue + newConstructionDummy*lotS
                 pctCollege*age + bathrooms*bedrooms, data = saratoga_train)

  #Testing
  yhat_test1 = predict(lm_medium, saratoga_test)
  yhat_test2 = predict(lm_handbuilt, saratoga_test)

  #Run it on the actual and the predicted values
  rmse1[i]= rmse(saratoga_test$price, yhat_test1)
  rmse2[i]= rmse(saratoga_test$price, yhat_test2)

}
```

```
## Warning in predict.lm(lm_handbuilt, saratoga_test): prediction from a rank-
## deficient fit may be misleading

## Warning in predict.lm(lm_handbuilt, saratoga_test): prediction from a rank-
## deficient fit may be misleading
```

```
mean(rmse1)
```

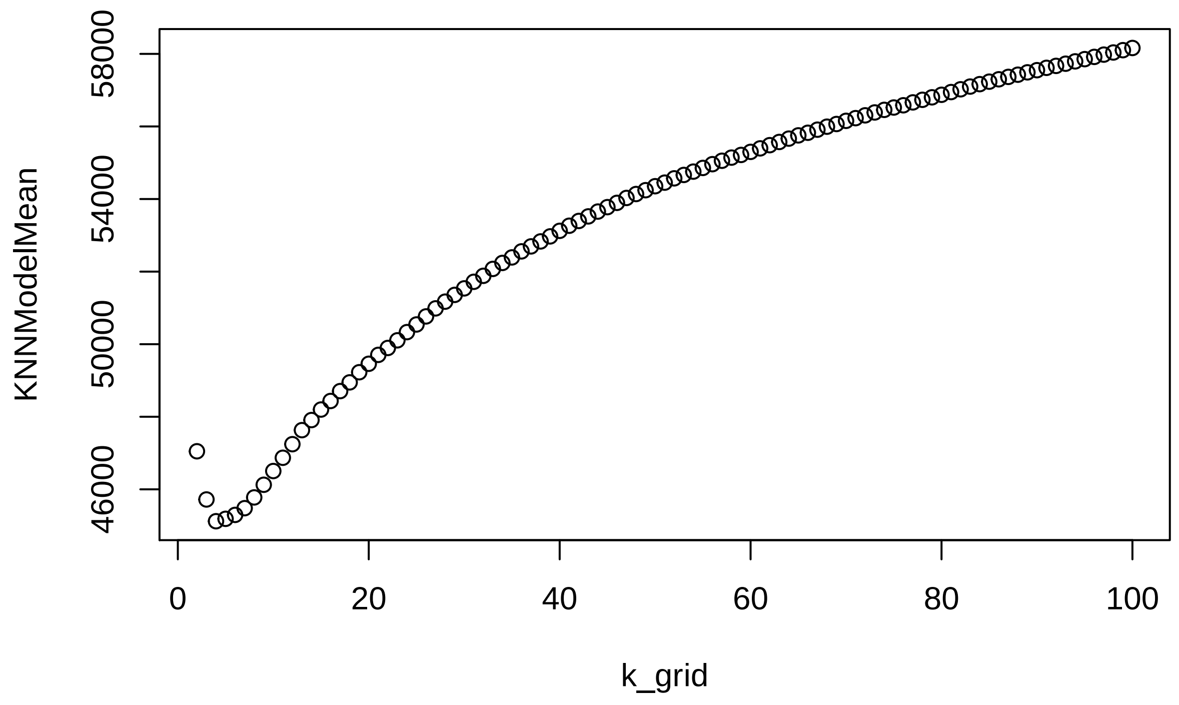
```
## [1] 61176.54
```

```
mean(rmse2)
```

```
## [1] 3.091566e-11
```

Attempt at turning my hand-built linear model into a better-performing KNN model:

```
# K-Nearest Neighbors Model

#Defining train-test sets for the hand-built regression model

KNNModel = do(100)*{
  N = nrow(SaratogaHouses)
  train = round(0.8*N)
  test = (N-train)

  train_cases = sample.int(N, train, replace=FALSE)
  test_cases = setdiff(1:N, train_cases)

  saratoga_train = SaratogaHouses[train_cases,]
  saratoga_test = SaratogaHouses[test_cases,]

  Xtrain = model.matrix(~ lotSize + age + livingArea + pctCollege + bedrooms +
                 fireplaces + bathrooms + rooms + heating + fuel + centralAir + ConstructionCost
                 - 1, data=saratoga_train)
  Xtest = model.matrix(~ lotSize + age + livingArea + pctCollege + bedrooms +
                 fireplaces + bathrooms + rooms + heating + fuel + centralAir + ConstructionCost
                 - 1, data=saratoga_test)

  Ytrain = saratoga_train$price
  Ytest = saratoga_test$price
  #Scaling the features (Standardization)
  scale_train = apply(Xtrain, 2, sd)
  Xtilde_train = scale(Xtrain, scale = scale_train)
  Xtilde_test = scale(Xtest, scale = scale_train)

  #The for loop
    k_grid = seq(2,100)
    rmse_grid = foreach(K = k_grid, .combine='c') %do% {
      KNNModel = knn.reg(Xtilde_train, Xtilde_test, Ytrain, k=K)
    rmse(Ytest, KNNModel$pred)
  }
}
KNNModelMean = colMeans(KNNModel)

#Plotting
plot(k_grid, KNNModelMean)
abline(h=rmse(Ytest, yhat_test2))
```



Conclusion:

While building our model we realized that when a variable does not completely capture all the information about the house then we should not eliminate it without giving it any thought because then we may lose some important information. For example, we should not eliminate bathrooms and bedrooms variables because knowing how many bathrooms and bedrooms specifically is important for buyers which is not fully captured by the rooms variable. On the other hand, we cannot eliminate rooms and only have bedrooms and bathrooms because bedrooms and bathrooms are not the only type of rooms that affects house prices.

Other types of rooms such as laundry room, storeroom, sunroom, etc. are also included in rooms and how many rooms besides bathrooms and bedrooms are important in determining house prices. We have added one composite variable ConstructionCost and five interaction terms to the medium model to improve the model's predictive performance.The variable ConstructionCost and the interactions of ConstructionCost and landValue, newConstructionDummy and landValue, newConstructionDummy and lotSize, pctCollege and age, and bathrooms and bedrooms, all seem to be especially strong drivers of house prices because addicting these has decreased the rmse from something in the 60,000 to almost 0 depending on random variation.

ConstructionCost is calculated by subtracting landvaule from price and used to represent the cost of building the house on any piece of land. The interaction of ConstructionCost and landvalue is capturing how luxurious houses using high-quality raw materials are built in neighborhoods where land value is high or appreciating. Similarly cheaper houses are built in neighborhoods where land value is low because poorer people live there. The interaction term newConstructionDummy and landvlue is capturing that new constructions happened in neighborhoods where land value is higher or appreciating and the interaction of newConstructionDummy and lotSize is capturing the fact that more new construction happens as lot size increases. The interaction of pctCollege and age captures the fact that the older the houses are, the higher is the percent of the neighborhood that graduated from college.The interaction of bedroom and bathroom captures that the more bedrooms a house the more bathrooms it will have and so the higher will be the price of the house.

From our analysis, it appears that your organization can tax newly built houses slightly more because after analyzing the data we have found that newer houses are bigger and are correlated with an increase in pricing. Normally these houses most probably belong to the rich class of society. We will not recommend you to charge newer built houses a really high rate though because it also appears that age of the house is correlated with the percentage of college graduates living in the neighborhood and the higher the age and/or percent of college graduates, the higher is the price of the house. In most cases, the former effect of newly built houses dominates the latter effect of age and percent of college graduates in the neighborhood.