Homework 1

Yufei Lin, Jingfeng Xia

Nov 7 2020

```
#install.packages('pls')
#install.packages('randomForest')
library(pls)
```

```
##
## Attaching package: 'pls'

## The following object is masked from 'package:stats':
##
##     loadings
```

```
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(gam)
```

```
## Loading required package: splines

## Loading required package: foreach

##
## Attaching package: 'foreach'

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

## Loaded gam 1.16.1
```

Data Processing

Read in Data

```
curDir = getwd()
setwd(curDir)
vault = read.csv("./SourceData/test.csv")
HousePricing = read.csv("./SourceData/train.csv")
```

Seperate into Test and Training Set

```r
set.seed(1)
randS = sample(1:nrow(HousePricing), nrow(HousePricing)/2)
train = HousePricing[randS,]
test = HousePricing[-randS,]
summary(test)
```

```
##        Id           MSSubClass        MSZoning    LotFrontage
##  Min.   :   1.0   Min.   : 20.00   C (all):  8   Min.   :  0.00
##  1st Qu.: 396.2   1st Qu.: 20.00   FV     : 32   1st Qu.: 44.00
##  Median : 776.5   Median : 50.00   RH     : 10   Median : 64.00
##  Mean   : 755.1   Mean   : 58.15   RL     :574   Mean   : 58.96
##  3rd Qu.:1119.8   3rd Qu.: 75.00   RM     :106   3rd Qu.: 79.00
##  Max.   :1460.0   Max.   :190.00                 Max.   :313.00
##
##    LotArea        Street     Alley     LotShape  LandContour  Utilities
##  Min.   :  1300   Grvl:  6   0   :689   IR1:224   Bnk: 28     AllPub:729
##  1st Qu.:  7500   Pave:724   Grvl: 22   IR2: 21   HLS: 30     NoSeWa:  1
##  Median :  9552              Pave: 19   IR3:  5   Low: 20
##  Mean   : 10585                         Reg:480   Lvl:652
##  3rd Qu.: 11439
##  Max.   :215245
##
##    LotConfig   LandSlope  Neighborhood   Condition1    Condition2    BldgType
##  Corner :132   Gtl:686   Omes   :113   Norm   :638   Norm   :724   1Fam  :595
##  CulDSac: 35   Mod: 37   CollgCr: 72   Feedr  : 38   Artery :  2   2fmCon: 15
##  FR2    : 28   Sev:  7   OldTown: 54   Artery : 26   Feedr  :  1   Duplex: 32
##  FR3    :  0             NridgHt: 48   RRAn   : 10   PosA   :  1   Twnhs : 25
##  Inside :535             Edwards: 45   RRAe   :  6   PosN   :  1   TwnhsE: 63
##                          Somerst: 43   PosN   :  5   RRAn   :  1
##                          (Other):355   (Other):  7   (Other):  0
##    HouseStyle   OverallQual      OverallCond      YearBuilt      YearRemodAdd
##  1Story :374   Min.   : 1.000   Min.   :2.000   Min.   :1880   Min.   :1950
##  2Story :213   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967
##  1.5Fin : 75   Median : 6.000   Median :5.000   Median :1975   Median :1994
##  SLvl   : 37   Mean   : 6.096   Mean   :5.568   Mean   :1972   Mean   :1985
##  SFoyer : 17   3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2001   3rd Qu.:2004
##  1.5Unf :  6   Max.   :10.000   Max.   :9.000   Max.   :2010   Max.   :2010
##  (Other):  8
##    RoofStyle      RoofMatl    Exterior1st   Exterior2nd    MasVnrType
##  Flat   :  6   CompShg:719   VinylSd:253   VinylSd:246   0      :  4
##  Gable  :567   Tar&Grv:  5   MetalSd:121   MetalSd:117   BrkCmn :  5
##  Gambrel:  6   WdShngl:  3   HdBoard:117   HdBoard:105   BrkFace:225
##  Hip    :148   WdShake:  2   Wd Sdng: 86   Wd Sdng: 90   None   :425
##  Mansard:  2   Membran:  1   Plywood: 59   Plywood: 75   Stone  : 71
##  Shed   :  1   ClyTile:  0   CemntBd: 31   CmentBd: 31
##                (Other):  0   (Other): 63   (Other): 66
##    MasVnrArea     ExterQual ExterCond  Foundation  BsmtQual BsmtCond
##  Min.   :   0.0   Ex: 30    Ex:  3    BrkTil: 71   0 : 20   0 : 20
##  1st Qu.:   0.0   Fa:  7    Fa: 11    CBlock:318   Ex: 63   Fa: 21
##  Median :   0.0   Gd:236    Gd: 64    PConc :323   Fa: 17   Gd: 38
##  Mean   : 108.4   TA:457    Po:  1    Slab  : 14   Gd:309   Po:  0
##  3rd Qu.: 168.8             TA:651    Stone :  2   TA:321   TA:651
##  Max.   :1378.0                       Wood  :  2
##
##  BsmtExposure BsmtFinType1  BsmtFinSF1     BsmtFinType2  BsmtFinSF2
##  0 : 21       0  : 20      Min.   :   0.0   0  : 21      Min.   :   0.00
##  Av: 99       ALQ:114      1st Qu.:   0.0   ALQ:  9      1st Qu.:   0.00
##  Gd: 71       BLQ: 72      Median : 401.0   BLQ: 18      Median :   0.00
##  Mn: 56       GLQ:224      Mean   : 460.5   GLQ:  7      Mean   :  45.61
##  No:483       LwQ: 35      3rd Qu.: 738.8   LwQ: 23      3rd Qu.:   0.00
##               Rec: 61      Max.   :2188.0   Rec: 28      Max.   :1080.00
##               Unf:204                       Unf:624
##    BsmtUnfSF       TotalBsmtSF    Heating    HeatingQC CentralAir Electrical
##  Min.   :   0.0   Min.   :   0   Floor:  1   Ex:362    N: 39      0    :  0
##  1st Qu.: 203.2   1st Qu.: 780   GasA :714   Fa: 23    Y:691      FuseA: 38
##  Median : 441.5   Median : 990   GasW : 10   Gd:124               FuseF: 11
##  Mean   : 547.0   Mean   :1053   Grav :  3   Po:  0               FuseP:  1
##  3rd Qu.: 775.0   3rd Qu.:1318   OthW :  0   TA:221               Mix  :  0
##  Max.   :2121.0   Max.   :3206   Wall :  2                        SBrkr:680
##
##    X1stFlrSF      X2ndFlrSF       LowQualFinSF       GrLivArea
##  Min.   : 334   Min.   :   0.0   Min.   :  0.000   Min.   : 334
##  1st Qu.: 876   1st Qu.:   0.0   1st Qu.:  0.000   1st Qu.:1124
##  Median :1081   Median :   0.0   Median :  0.000   Median :1456
##  Mean   :1161   Mean   : 330.5   Mean   :  4.197   Mean   :1495
##  3rd Qu.:1392   3rd Qu.: 708.8   3rd Qu.:  0.000   3rd Qu.:1740
##  Max.   :3228   Max.   :1872.0   Max.   :572.000   Max.   :4316
##
##   BsmtFullBath    BsmtHalfBath       FullBath       HalfBath
##  Min.   :0.000   Min.   :0.00000   Min.   :0.00   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:1.00   1st Qu.:0.0000
##  Median :0.000   Median :0.00000   Median :2.00   Median :0.0000
##  Mean   :0.437   Mean   :0.06301   Mean   :1.57   Mean   :0.3685
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:2.00   3rd Qu.:1.0000
##  Max.   :2.000   Max.   :1.00000   Max.   :3.00   Max.   :2.0000
##
##   BedroomAbvGr    KitcheObvGr    KitchenQual  TotRmsAbvGrd    Functio0l
##  Min.   :0.000   Min.   :1.000   Ex: 54      Min.   : 2.000   Maj1:  6
##  1st Qu.:2.000   1st Qu.:1.000   Fa: 19      1st Qu.: 5.000   Maj2:  3
##  Median :3.000   Median :1.000   Gd:287      Median : 6.000   Min1: 18
##  Mean   :2.875   Mean   :1.056   TA:370      Mean   : 6.493   Min2: 16
##  3rd Qu.:3.000   3rd Qu.:1.000               3rd Qu.: 7.000   Mod :  8
##  Max.   :6.000   Max.   :2.000               Max.   :12.000   Sev :  1
##                                                               Typ :678
##    Fireplaces     FireplaceQu   GarageType   GarageYrBlt   GarageFinish
##  Min.   :0.0000   0 :350      0       : 44   Min.   :   0   0  : 44
##  1st Qu.:0.0000   Ex: 12      2Types :  4   1st Qu.:1960   Fin:173
##  Median :1.0000   Fa: 13      Attchd :433   Median :1978   RFn:224
##  Mean   :0.6082   Gd:196      Basment: 12   Mean   :1860   Unf:289
##  3rd Qu.:1.0000   Po:  7      BuiltIn: 42   3rd Qu.:2001
##  Max.   :3.0000   TA:152      CarPort:  7   Max.   :2010
##                               Detchd :188
##    GarageCars      GarageArea     GarageQual GarageCond PavedDrive
##  Min.   :0.000   Min.   :   0.0   0 : 44     0 : 44     N: 40
##  1st Qu.:1.000   1st Qu.: 352.0   Ex:  2     Ex:  1     P: 16
##  Median :2.000   Median : 482.0   Fa: 23     Fa: 18     Y:674
##  Mean   :1.767   Mean   : 475.1   Gd:  8     Gd:  2
##  3rd Qu.:2.000   3rd Qu.: 576.0   Po:  0     Po:  2
##  Max.   :4.000   Max.   :1248.0   TA:653     TA:663
##
##    WoodDeckSF      OpenPorchSF     EnclosedPorch      X3SsnPorch
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :  0.00   Median : 22.00   Median :  0.00   Median :  0.000
##  Mean   : 94.17   Mean   : 45.33   Mean   : 21.79   Mean   :  2.886
##  3rd Qu.:169.50   3rd Qu.: 65.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :736.00   Max.   :547.00   Max.   :330.00   Max.   :304.000
##
##   ScreenPorch         PoolArea     PoolQC     Fence      MiscFeature
##  Min.   :  0.00   Min.   :  0.0   0 :728   0    :587   0   :700
##  1st Qu.:  0.00   1st Qu.:  0.0   Ex:  0   GdPrv: 25   Gar2:  1
##  Median :  0.00   Median :  0.0   Fa:  1   GdWo : 32   Othr:  1
##  Mean   : 13.79   Mean   :  1.5   Gd:  1   MnPrv: 82   Shed: 27
##  3rd Qu.:  0.00   3rd Qu.:  0.0            MnWw :  4   TenC:  1
##  Max.   :480.00   Max.   :576.0
##
##     MiscVal             MoSold           YrSold        SaleType    SaleCondition
##  Min.   :    0.00   Min.   : 1.000   Min.   :2006   WD     :629   Abnorml: 59
##  1st Qu.:    0.00   1st Qu.: 5.000   1st Qu.:2007   New    : 64   AdjLand:  0
##  Median :    0.00   Median : 6.000   Median :2008   COD    : 21   Alloca :  7
##  Mean   :   53.03   Mean   : 6.292   Mean   :2008   ConLD  :  5   Family :  7
##  3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009   ConLw  :  4   Normal :592
##  Max.   :15500.00   Max.   :12.000   Max.   :2010   ConLI  :  3   Partial: 65
##                                                     (Other):  4
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129600
##  Median :160600
##  Mean   :181329
##  3rd Qu.:215000
##  Max.   :755000
##
```

Feature Engineering

In this section, we convert all unnecessary NAs into suitable value based on the observation from data description on Kaggle website.

```
# Currently I have solved this by replacing all NA with 0 in excel,
# for some reason the replacement does not work in R.
```

Prediction Algorithms

1. PCA

```
set.seed(2)
# Need help with PCA analysis
#pcr.fit=pcr(SalePrice~., data=HousePricing, scale=TRUE, validation ="CV")
```

2. Random Forest

```
rfTrain=randomForest(SalePrice~.,data=train, mtry=6,importance =TRUE, na.action=na.roughfix)
rfYhat = predict(rfTrain ,newdata=test)
print("The test MSE is shown in the following: ")
```

```
## [1] "The test MSE is shown in the following: "
```

```
mean((rfYhat-test$SalePrice)^2)
```

```
## [1] 1066370931
```

```
temp = importance(rfTrain)
temp = as.data.frame(temp)
names(temp)[names(temp)=="%IncMSE"] <- "importance"
sort(temp$importance, decreasing = TRUE)
```

```
##  [1] 17.0165805 14.6396422 14.3804097 11.3432457 11.3252649 11.0803219
##  [7] 10.8442110  9.7307720  9.5903425  9.3398960  9.3228800  9.2399692
## [13]  9.0918306  8.6665417  8.0934299  7.9818180  7.7833588  7.6148368
## [19]  7.3383848  7.2505400  6.9147177  6.8963129  6.7623038  6.0894719
## [25]  6.0243481  5.8570122  5.8453338  5.7344228  4.9267802  4.9214221
## [31]  4.9179879  4.7575993  4.6577276  4.6461006  4.5045584  4.4889300
## [37]  4.4169840  4.3721631  4.3050122  4.1228497  4.1207557  4.0039098
## [43]  3.8947097  3.7686206  3.0789742  2.6792650  2.4509828  2.3792374
## [49]  2.3601484  2.1045215  1.9171621  1.8225149  1.6961293  1.6393663
## [55]  1.5393781  1.3906253  1.3447144  0.9266134  0.8786982  0.8407066
## [61]  0.8076229  0.7318058  0.6969641  0.5745699  0.4723516  0.1409855
## [67]  0.0000000  0.0000000 -0.1494808 -0.1675596 -0.5267200 -0.8747581
## [73] -1.1194777 -1.1224986 -1.1926411 -1.2034331 -1.4898769 -2.0869449
## [79] -2.6096805 -3.2285895
```

```
pander(temp, title="Importance of each factor according to random forest")
```

	importance	IncNodePurity
Id	-2.087	2.401e+10
MSSubClass	6.762	1.907e+10
MSZoning	7.615	2.048e+10
LotFrontage	-0.1495	3.972e+10
LotArea	7.338	9.639e+10
Street	0	0
Alley	0.5746	2.057e+09
LotShape	2.451	1.98e+10
LandContour	2.105	9.174e+09
Utilities	0	0
LotConfig	-1.119	1.015e+10

	importance	IncNodePurity
LandSlope	-0.8748	3.201e+09
Neighborhood	14.64	3.031e+11
Condition1	4.121	9.141e+09
Condition2	0.9266	1.306e+09
BldgType	4.505	6.65e+09
HouseStyle	6.915	2.94e+10
OverallQual	14.38	3.155e+11
OverallCond	7.251	1.55e+10
YearBuilt	9.34	1.598e+11
YearRemodAdd	8.667	7.126e+10
RoofStyle	-0.5267	1.076e+10
RoofMatl	-1.203	2.602e+09
Exterior1st	4.918	4.753e+10
Exterior2nd	4.921	5.516e+10
MasVnrType	4.004	2.52e+10
MasVnrArea	4.658	5.434e+10
ExterQual	11.33	2.116e+11
ExterCond	1.539	4.231e+09
Foundation	6.089	2.131e+10
BsmtQual	5.734	1.14e+11
BsmtCond	3.079	4.343e+09
BsmtExposure	0.8076	2.164e+10
BsmtFinType1	4.646	3.767e+10
BsmtFinSF1	6.024	8.662e+10
BsmtFinType2	1.696	1.049e+10
BsmtFinSF2	0.141	5.615e+09
BsmtUnfSF	4.372	3.077e+10
TotalBsmtSF	11.08	1.631e+11
Heating	-0.1676	1.257e+09
HeatingQC	4.927	2.502e+10
CentralAir	4.305	7.939e+09
Electrical	0.8407	2.798e+09
X1stFlrSF	9.731	1.387e+11
X2ndFlrSF	10.84	1.13e+11
LowQualFinSF	-1.193	848554186
GrLivArea	17.02	3.189e+11
BsmtFullBath	3.895	1.158e+10
BsmtHalfBath	-1.122	910528158
FullBath	6.896	1.051e+11
HalfBath	5.857	1.724e+10
BedroomAbvGr	4.758	2.489e+10
Kitche0bvGr	2.679	2.183e+09
KitchenQual	8.093	1.498e+11
TotRmsAbvGrd	5.845	8.734e+10
Functio0l	1.639	8.348e+09
Fireplaces	9.092	5.852e+10
FireplaceQu	9.59	8.616e+10
GarageType	7.783	4.958e+10
GarageYrBlt	9.24	1.209e+11
GarageFinish	7.982	9.176e+10
GarageCars	11.34	2.028e+11
GarageArea	9.323	2.147e+11

	importance	IncNodePurity
GarageQual	4.489	6.132e+09
GarageCond	3.769	7.115e+09
PavedDrive	1.917	3.812e+09
WoodDeckSF	4.123	2.822e+10
OpenPorchSF	4.417	4.994e+10
EnclosedPorch	0.697	6.085e+09
X3SsnPorch	1.391	626796305
ScreenPorch	1.345	6.633e+09
PoolArea	-1.49	2.243e+10
PoolQC	-2.61	2.605e+10
Fence	2.36	1.359e+10
MiscFeature	0.7318	582766091
MiscVal	0.8787	576558679
MoSold	0.4724	1.698e+10
YrSold	-3.229	9.169e+09
SaleType	2.379	1.618e+10
SaleCondition	1.823	1.833e+10

From the random forest analysis, we have discovered that the top three most important factors for predicting sale price are the following:

1. GrLivArea
2. Neighbourhood
3. OverallQual

Therefore, we will make GAM models according to these three factors.

```
fit1 = gam(SalePrice ~ GrLivArea + Neighborhood + OverallQual, data = HousePricing)
print("Deviance of Model 1")
```

```
## [1] "Deviance of Model 1"
```

```
deviance(fit1)
```

```
## [1] 1.962681e+12
```

From this we know that the deviance is quite large, we need a better model.

Cross Validation