# Ames Housing Feature Selection

*Emily Goren*

*4/6/2017*

## Data Cleaning

```
train <- read.csv(paste0(dir, "train.csv"))
train <- subset(train, select = -Id) # Drop ID, useless for prediction.
dim(train)
```

```
## [1] 1460   80
```

```
str(train)
```

```
## 'data.frame':    1460 obs. of  80 variables:
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
##  $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA NA ...
##  $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1 4 4 ...
##  $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...
##  $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...
##  $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...
##  $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...
##  $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...
##  $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
##  $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...
##  $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...
##  $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4 4 ...
##  $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
##  $ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4 4 ...
##  $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4 4 ...
##  $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...
##  $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 6 3 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2 6 6 ...
##  $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
```

```
##  $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
##  $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3 1 ...
##  $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...
##  $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
##  $ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7 ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5 5 5 ...
##  $ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
##  $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2 3 ...
##  $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...
##  $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 NA NA NA NA ...
##  $ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA 3 NA NA ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
##  $ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

```
summary(train)
```

```
##    MSSubClass       MSZoning     LotFrontage        LotArea
##  Min.   : 20.0   C (all):  10   Min.   : 21.00   Min.   :  1300
##  1st Qu.: 20.0   FV     :  65   1st Qu.: 59.00   1st Qu.:  7554
##  Median : 50.0   RH     :  16   Median : 69.00   Median :  9478
##  Mean   : 56.9   RL     :1151   Mean   : 70.05   Mean   : 10517
##  3rd Qu.: 70.0   RM     : 218   3rd Qu.: 80.00   3rd Qu.: 11602
##  Max.   :190.0                  Max.   :313.00   Max.   :215245
##                                 NA's   :259
```

```
##   Street      Alley      LotShape  LandContour  Utilities
##  Grvl:   6   Grvl:  50   IR1:484   Bnk:  63    AllPub:1459
##  Pave:1454   Pave:  41   IR2: 41   HLS:  50    NoSeWa:   1
##              NA's:1369   IR3: 10   Low:  36
##                          Reg:925   Lvl:1311
##
##
##
##    LotConfig    LandSlope   Neighborhood   Condition1     Condition2
##  Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260   Norm   :1445
##  CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81   Feedr  :   6
##  FR2    :  47   Sev:  13   OldTown:113   Artery :  48   Artery :   2
##  FR3    :   4              Edwards:100   RRAn   :  26   PosN   :   2
##  Inside :1052              Somerst: 86   PosN   :  19   RRNn   :   2
##                            Gilbert: 79   RRAe   :  11   PosA   :   1
##                            (Other):707   (Other):  15   (Other):   2
##    BldgType      HouseStyle   OverallQual      OverallCond
##  1Fam  :1220   1Story :726   Min.   : 1.000   Min.   :1.000
##  2fmCon:  31   2Story :445   1st Qu.: 5.000   1st Qu.:5.000
##  Duplex:  52   1.5Fin :154   Median : 6.000   Median :5.000
##  Twnhs :  43   SLvl   : 65   Mean   : 6.099   Mean   :5.575
##  TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000   3rd Qu.:6.000
##                1.5Unf : 14   Max.   :10.000   Max.   :9.000
##                (Other): 19
##    YearBuilt     YearRemodAdd    RoofStyle       RoofMatl     Exterior1st
##  Min.   :1872   Min.   :1950   Flat   :  13   CompShg:1434   VinylSd:515
##  1st Qu.:1954   1st Qu.:1967   Gable  :1141   Tar&Grv:  11   HdBoard:222
##  Median :1973   Median :1994   Gambrel:  11   WdShngl:   6   MetalSd:220
##  Mean   :1971   Mean   :1985   Hip    : 286   WdShake:   5   Wd Sdng:206
##  3rd Qu.:2000   3rd Qu.:2004   Mansard:   7   ClyTile:   1   Plywood:108
##  Max.   :2010   Max.   :2010   Shed   :   2   Membran:   1   CemntBd: 61
##                                               (Other):   2   (Other):128
##   Exterior2nd    MasVnrType    MasVnrArea     ExterQual ExterCond
##  VinylSd:504   BrkCmn : 15   Min.   :   0.0   Ex: 52    Ex:   3
##  MetalSd:214   BrkFace:445   1st Qu.:   0.0   Fa: 14    Fa:  28
##  HdBoard:207   None   :864   Median :   0.0   Gd:488    Gd: 146
##  Wd Sdng:197   Stone  :128   Mean   : 103.7   TA:906    Po:   1
##  Plywood:142   NA's   :  8   3rd Qu.: 166.0             TA:1282
##  CmentBd: 60                 Max.   :1600.0
##  (Other):136                 NA's   :8
##   Foundation  BsmtQual   BsmtCond    BsmtExposure BsmtFinType1
##  BrkTil:146   Ex  :121   Fa  :  45   Av  :221     ALQ :220
##  CBlock:634   Fa  : 35   Gd  :  65   Gd  :134     BLQ :148
##  PConc :647   Gd  :618   Po  :   2   Mn  :114     GLQ :418
##  Slab  : 24   TA  :649   TA  :1311   No  :953     LwQ : 74
##  Stone :  6   NA's: 37   NA's:  37   NA's: 38     Rec :133
##  Wood  :  3                                      Unf :430
##                                                  NA's: 37
##    BsmtFinSF1     BsmtFinType2   BsmtFinSF2       BsmtUnfSF
##  Min.   :   0.0   ALQ :  19    Min.   :   0.00   Min.   :   0.0
##  1st Qu.:   0.0   BLQ :  33    1st Qu.:   0.00   1st Qu.: 223.0
##  Median : 383.5   GLQ :  14    Median :   0.00   Median : 477.5
##  Mean   : 443.6   LwQ :  46    Mean   :  46.55   Mean   : 567.2
##  3rd Qu.: 712.2   Rec :  54    3rd Qu.:   0.00   3rd Qu.: 808.0
```

```
##  Max.   :5644.0   Unf :1256    Max.   :1474.00   Max.   :2336.0
##                   NA's:  38
##   TotalBsmtSF      Heating     HeatingQC CentralAir Electrical
##  Min.   :   0.0   Floor:   1   Ex:741    N:  95     FuseA:  94
##  1st Qu.: 795.8   GasA :1428   Fa: 49    Y:1365     FuseF:  27
##  Median : 991.5   GasW :  18   Gd:241               FuseP:   3
##  Mean   :1057.4   Grav :   7   Po:  1               Mix  :   1
##  3rd Qu.:1298.2   OthW :   2   TA:428               SBrkr:1334
##  Max.   :6110.0   Wall :   4                        NA's :   1
##
##    X1stFlrSF      X2ndFlrSF     LowQualFinSF       GrLivArea
##  Min.   : 334   Min.   :   0   Min.   :  0.000   Min.   : 334
##  1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130
##  Median :1087   Median :   0   Median :  0.000   Median :1464
##  Mean   :1163   Mean   : 347   Mean   :  5.845   Mean   :1515
##  3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777
##  Max.   :4692   Max.   :2065   Max.   :572.000   Max.   :5642
##
##   BsmtFullBath     BsmtHalfBath        FullBath        HalfBath
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000   Median :2.000   Median :0.0000
##  Mean   :0.4253   Mean   :0.05753   Mean   :1.565   Mean   :0.3829
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :3.0000   Max.   :2.00000   Max.   :3.000   Max.   :2.0000
##
##   BedroomAbvGr    KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional
##  Min.   :0.000   Min.   :0.000   Ex:100      Min.   : 2.000   Maj1:  14
##  1st Qu.:2.000   1st Qu.:1.000   Fa: 39      1st Qu.: 5.000   Maj2:   5
##  Median :3.000   Median :1.000   Gd:586      Median : 6.000   Min1:  31
##  Mean   :2.866   Mean   :1.047   TA:735      Mean   : 6.518   Min2:  34
##  3rd Qu.:3.000   3rd Qu.:1.000               3rd Qu.: 7.000   Mod :  15
##  Max.   :8.000   Max.   :3.000               Max.   :14.000   Sev :   1
##                                                               Typ :1360
##    Fireplaces    FireplaceQu   GarageType   GarageYrBlt   GarageFinish
##  Min.   :0.000   Ex  : 24    2Types :  6   Min.   :1900   Fin :352
##  1st Qu.:0.000   Fa  : 33    Attchd :870   1st Qu.:1961   RFn :422
##  Median :1.000   Gd  :380    Basment: 19   Median :1980   Unf :605
##  Mean   :0.613   Po  : 20    BuiltIn: 88   Mean   :1979   NA's: 81
##  3rd Qu.:1.000   TA  :313    CarPort:  9   3rd Qu.:2002
##  Max.   :3.000   NA's:690    Detchd :387   Max.   :2010
##                              NA's   : 81   NA's   :81
##    GarageCars      GarageArea     GarageQual  GarageCond  PavedDrive
##  Min.   :0.000   Min.   :   0.0   Ex  :   3   Ex  :   2   N:  90
##  1st Qu.:1.000   1st Qu.: 334.5   Fa  :  48   Fa  :  35   P:  30
##  Median :2.000   Median : 480.0   Gd  :  14   Gd  :   9   Y:1340
##  Mean   :1.767   Mean   : 473.0   Po  :   3   Po  :   7
##  3rd Qu.:2.000   3rd Qu.: 576.0   TA  :1311   TA  :1326
##  Max.   :4.000   Max.   :1418.0   NA's:  81   NA's:  81
##
##    WoodDeckSF      OpenPorchSF     EnclosedPorch      X3SsnPorch
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Median :  0.00   Median : 25.00   Median :  0.00   Median :  0.00
```

```
##  Mean   : 94.24   Mean   : 46.66   Mean   : 21.95   Mean   :  3.41
##  3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00   3rd Qu.:  0.00
##  Max.   :857.00   Max.   :547.00   Max.   :552.00   Max.   :508.00
##
##   ScreenPorch        PoolArea        PoolQC       Fence      MiscFeature
##  Min.   :  0.00   Min.   :  0.000   Ex  :   2   GdPrv:  59   Gar2:   2
##  1st Qu.:  0.00   1st Qu.:  0.000   Fa  :   2   GdWo :  54   Othr:   2
##  Median :  0.00   Median :  0.000   Gd  :   3   MnPrv: 157   Shed:  49
##  Mean   : 15.06   Mean   :  2.759   NA's:1453   MnWw :  11   TenC:   1
##  3rd Qu.:  0.00   3rd Qu.:  0.000               NA's :1179   NA's:1406
##  Max.   :480.00   Max.   :738.000
##
##     MiscVal             MoSold           YrSold        SaleType
##  Min.   :    0.00   Min.   : 1.000   Min.   :2006   WD     :1267
##  1st Qu.:    0.00   1st Qu.: 5.000   1st Qu.:2007   New    : 122
##  Median :    0.00   Median : 6.000   Median :2008   COD    :  43
##  Mean   :   43.49   Mean   : 6.322   Mean   :2008   ConLD  :   9
##  3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009   ConLI  :   5
##  Max.   :15500.00   Max.   :12.000   Max.   :2010   ConLw  :   5
##                                                     (Other):   9
##  SaleCondition    SalePrice
##  Abnorml: 101   Min.   : 34900
##  AdjLand:   4   1st Qu.:129975
##  Alloca :  12   Median :163000
##  Family :  20   Mean   :180921
##  Normal :1198   3rd Qu.:214000
##  Partial: 125   Max.   :755000
##
```

Building on Andrew's observations:

- Change MSSubClass to a factor.
- For numeric variables, replace `NA`'s with the median value.
- For factor variables, make NA's into their own factor level. Most of these are "not applicable"" so they'll likely be correlated with similar variables, e.g. garage area and garage quality.
- Change the two condition variables into binary indicators for each condition.

```
train$MSSubClass <- as.factor(train$MSSubClass)
# Replace NA's.
missing <- apply(train, 2, function(x) sum(is.na(x)))
missing[missing > 0]
```

```
##  LotFrontage        Alley   MasVnrType   MasVnrArea     BsmtQual
##          259         1369            8            8           37
##     BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2   Electrical
##           37           38           37           38            1
##  FireplaceQu   GarageType  GarageYrBlt GarageFinish   GarageQual
##          690           81           81           81           81
##   GarageCond       PoolQC        Fence  MiscFeature
##           81         1453         1179         1406
```

```
missvars <- names(missing[missing > 0])
str(subset(train, select = missvars))
```

```
## 'data.frame':    1460 obs. of  19 variables:
##  $ LotFrontage : int  65 80 68 60 84 85 75 NA 51 50 ...
```

```
##  $ Alley       : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA NA ...
##  $ MasVnrType  : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...
##  $ MasVnrArea  : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ BsmtQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4 4 ...
##  $ BsmtCond    : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4 4 ...
##  $ BsmtExposure: Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...
##  $ BsmtFinType1: Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 6 3 ...
##  $ BsmtFinType2: Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2 6 6 ...
##  $ Electrical  : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...
##  $ FireplaceQu : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5 5 5 ...
##  $ GarageType  : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...
##  $ GarageYrBlt : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish: Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
##  $ GarageQual  : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2 3 ...
##  $ GarageCond  : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ PoolQC      : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...
##  $ Fence       : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 NA NA NA NA ...
##  $ MiscFeature : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA 3 NA NA ...
```

```
# Replace numeric missing values with median, add NA as a factor level otherwise
for (i in 1:length(missvars)) {
    thisvar <- train[ , missvars[i]]
    if (is.numeric(thisvar))
        thisvar[is.na(thisvar)] <-  median(thisvar, na.rm = TRUE)
    if (is.factor(thisvar))
        thisvar <- addNA(thisvar)
    train[ , missvars[i]] <- thisvar
}
# Make indicators for conditions.
cond1 <- data.frame(model.matrix(~ Condition1 + 0, data = train))
names(cond1) <- sub(".*1", "", names(cond1))
cond2 <- data.frame(model.matrix(~ Condition2 + 0, data = train))
names(cond2) <- sub(".*2", "", names(cond2))
idx <- names(cond1) %in% names(cond2)
cond <- cond1
cond[, idx] <- cond1[, idx] + cond2
cond <- as.data.frame(ifelse(cond == 0, 0, 1))
train <- subset(train, select = -c(Condition1, Condition2))
train <- cbind(train, cond)
```

# Feature Importance in (untuned) Models

## Random Forest

```
fit.rf <- randomForest(SalePrice ~ ., data = train, importance = TRUE)
imp.rf <- varImp(fit.rf)
```

## Linear Model

```
fit.lm <- lm(SalePrice ~ ., data = train)
imp.lm <- varImp(fit.lm)
```

### Boosted Tree

```
X <- subset(train, select = -SalePrice)
fit.bst <- xgboost(data.matrix(X), train$SalePrice, nrounds = 200)
```

```
## [1]  train-rmse:141321.500000
## [2]  train-rmse:101740.679688
## [3]  train-rmse:73956.531250
## [4]  train-rmse:54460.515625
## [5]  train-rmse:40688.312500
## [6]  train-rmse:31187.535156
## [7]  train-rmse:24619.798828
## [8]  train-rmse:20087.957031
## [9]  train-rmse:16968.443359
## [10] train-rmse:14754.686523
## [11] train-rmse:13201.319336
## [12] train-rmse:12275.466797
## [13] train-rmse:11561.325195
## [14] train-rmse:11072.065430
## [15] train-rmse:10587.292969
## [16] train-rmse:10177.582031
## [17] train-rmse:9883.966797
## [18] train-rmse:9557.189453
## [19] train-rmse:9358.337891
## [20] train-rmse:9144.688477
## [21] train-rmse:8872.632812
## [22] train-rmse:8717.582031
## [23] train-rmse:8498.416992
## [24] train-rmse:8288.791016
## [25] train-rmse:7923.717773
## [26] train-rmse:7781.065918
## [27] train-rmse:7605.622559
## [28] train-rmse:7524.075195
## [29] train-rmse:7218.084473
## [30] train-rmse:7039.415527
## [31] train-rmse:6781.976562
## [32] train-rmse:6730.637207
## [33] train-rmse:6630.716797
## [34] train-rmse:6522.078125
## [35] train-rmse:6397.410645
## [36] train-rmse:6254.017090
## [37] train-rmse:6132.418945
## [38] train-rmse:5943.267090
## [39] train-rmse:5867.226074
## [40] train-rmse:5577.573730
## [41] train-rmse:5535.217285
## [42] train-rmse:5381.810059
## [43] train-rmse:5156.446289
## [44] train-rmse:5142.315430
```

```
## [45] train-rmse:5064.861328
## [46] train-rmse:4993.667969
## [47] train-rmse:4903.521484
## [48] train-rmse:4823.670410
## [49] train-rmse:4778.952148
## [50] train-rmse:4724.563965
## [51] train-rmse:4618.430664
## [52] train-rmse:4497.829102
## [53] train-rmse:4448.106934
## [54] train-rmse:4309.299316
## [55] train-rmse:4233.343750
## [56] train-rmse:4136.234863
## [57] train-rmse:4087.788574
## [58] train-rmse:3985.692871
## [59] train-rmse:3948.410889
## [60] train-rmse:3833.608887
## [61] train-rmse:3739.167725
## [62] train-rmse:3640.073730
## [63] train-rmse:3560.270264
## [64] train-rmse:3515.912354
## [65] train-rmse:3390.345947
## [66] train-rmse:3318.757568
## [67] train-rmse:3235.859131
## [68] train-rmse:3225.224854
## [69] train-rmse:3200.210693
## [70] train-rmse:3119.243408
## [71] train-rmse:3045.248779
## [72] train-rmse:2960.706787
## [73] train-rmse:2943.895996
## [74] train-rmse:2851.930420
## [75] train-rmse:2722.457764
## [76] train-rmse:2695.894287
## [77] train-rmse:2624.376709
## [78] train-rmse:2603.578613
## [79] train-rmse:2529.829346
## [80] train-rmse:2449.067871
## [81] train-rmse:2353.206543
## [82] train-rmse:2341.180420
## [83] train-rmse:2331.271973
## [84] train-rmse:2285.537598
## [85] train-rmse:2256.982178
## [86] train-rmse:2229.165039
## [87] train-rmse:2176.491699
## [88] train-rmse:2081.509033
## [89] train-rmse:2031.449341
## [90] train-rmse:2009.952637
## [91] train-rmse:1989.810913
## [92] train-rmse:1958.950317
## [93] train-rmse:1929.692261
## [94] train-rmse:1907.803589
## [95] train-rmse:1870.793213
## [96] train-rmse:1859.907715
## [97] train-rmse:1824.732178
## [98] train-rmse:1805.246338
```

```
## [99] train-rmse:1791.834351
## [100]    train-rmse:1755.693481
## [101]    train-rmse:1696.510742
## [102]    train-rmse:1652.962524
## [103]    train-rmse:1595.359253
## [104]    train-rmse:1589.844971
## [105]    train-rmse:1561.526123
## [106]    train-rmse:1554.139526
## [107]    train-rmse:1541.367798
## [108]    train-rmse:1512.762573
## [109]    train-rmse:1486.345215
## [110]    train-rmse:1480.471436
## [111]    train-rmse:1454.066406
## [112]    train-rmse:1437.221191
## [113]    train-rmse:1425.156128
## [114]    train-rmse:1396.535400
## [115]    train-rmse:1351.799438
## [116]    train-rmse:1329.628418
## [117]    train-rmse:1294.475952
## [118]    train-rmse:1264.744629
## [119]    train-rmse:1258.728394
## [120]    train-rmse:1235.047485
## [121]    train-rmse:1207.694336
## [122]    train-rmse:1171.680298
## [123]    train-rmse:1160.328735
## [124]    train-rmse:1127.340576
## [125]    train-rmse:1123.284546
## [126]    train-rmse:1108.357544
## [127]    train-rmse:1097.551392
## [128]    train-rmse:1084.931641
## [129]    train-rmse:1062.288940
## [130]    train-rmse:1039.389038
## [131]    train-rmse:1025.626221
## [132]    train-rmse:1006.930725
## [133]    train-rmse:980.808838
## [134]    train-rmse:955.087646
## [135]    train-rmse:948.012695
## [136]    train-rmse:927.379822
## [137]    train-rmse:915.200134
## [138]    train-rmse:903.427795
## [139]    train-rmse:878.080872
## [140]    train-rmse:871.130188
## [141]    train-rmse:860.971252
## [142]    train-rmse:845.894043
## [143]    train-rmse:817.361023
## [144]    train-rmse:795.785278
## [145]    train-rmse:772.884705
## [146]    train-rmse:755.341797
## [147]    train-rmse:741.355835
## [148]    train-rmse:732.914612
## [149]    train-rmse:725.699829
## [150]    train-rmse:698.567078
## [151]    train-rmse:687.887207
## [152]    train-rmse:673.794312
```

```
## [153]    train-rmse:670.559082
## [154]    train-rmse:659.723877
## [155]    train-rmse:650.448303
## [156]    train-rmse:636.126038
## [157]    train-rmse:628.581482
## [158]    train-rmse:623.267090
## [159]    train-rmse:604.202026
## [160]    train-rmse:575.603455
## [161]    train-rmse:572.943359
## [162]    train-rmse:562.816223
## [163]    train-rmse:545.429443
## [164]    train-rmse:542.976746
## [165]    train-rmse:538.681458
## [166]    train-rmse:518.515503
## [167]    train-rmse:502.110352
## [168]    train-rmse:494.294891
## [169]    train-rmse:485.999817
## [170]    train-rmse:479.843933
## [171]    train-rmse:473.485504
## [172]    train-rmse:463.650726
## [173]    train-rmse:448.735870
## [174]    train-rmse:434.851501
## [175]    train-rmse:424.264435
## [176]    train-rmse:415.776367
## [177]    train-rmse:411.085419
## [178]    train-rmse:405.694489
## [179]    train-rmse:399.748199
## [180]    train-rmse:379.921997
## [181]    train-rmse:372.964569
## [182]    train-rmse:363.992615
## [183]    train-rmse:358.033661
## [184]    train-rmse:348.836548
## [185]    train-rmse:343.695618
## [186]    train-rmse:338.225189
## [187]    train-rmse:335.291260
## [188]    train-rmse:332.339233
## [189]    train-rmse:325.413818
## [190]    train-rmse:319.284363
## [191]    train-rmse:317.031128
## [192]    train-rmse:312.907410
## [193]    train-rmse:301.075531
## [194]    train-rmse:295.178070
## [195]    train-rmse:287.267487
## [196]    train-rmse:284.801666
## [197]    train-rmse:281.540070
## [198]    train-rmse:274.298187
## [199]    train-rmse:271.474335
## [200]    train-rmse:266.002228
```

```
imp.bst <- xgb.importance(model = fit.bst, feature_names = names(X))
```

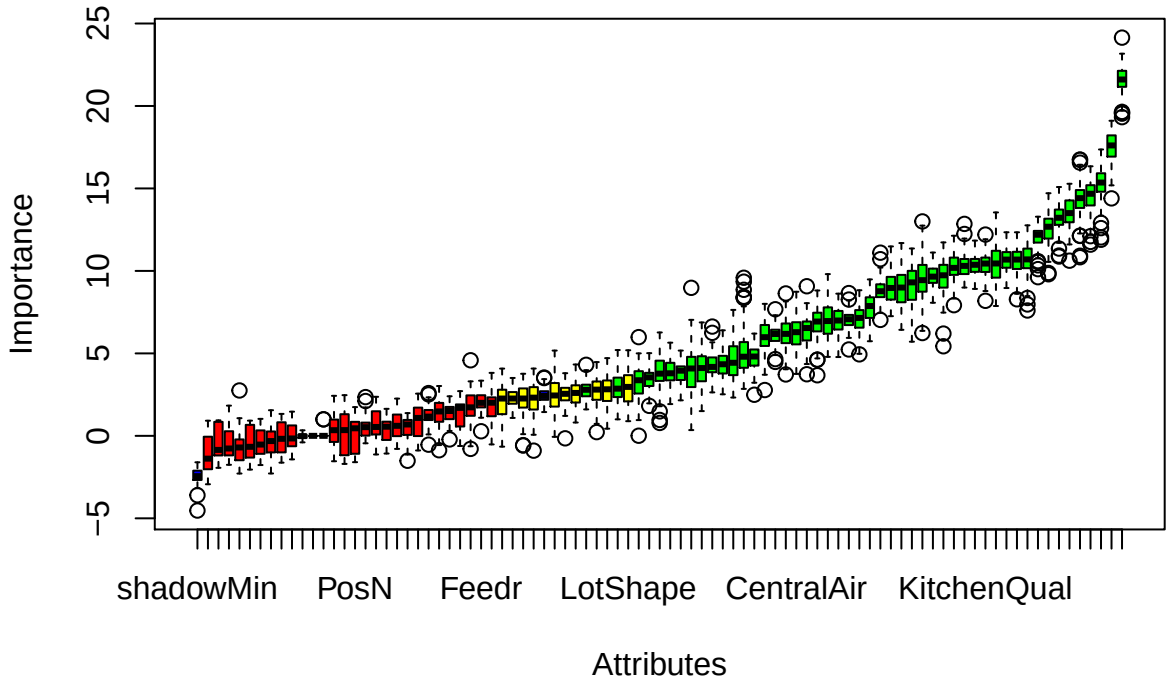
### Boruta

A feature selection method using random forests (see Kursa & Rudnicki, J Stat Software, 2010).

```
fit.bt <- Boruta(SalePrice ~ ., data = train)
print(fit.bt)
```

```
## Boruta performed 99 iterations in 2.762414 mins.
##  49 attributes confirmed important: BedroomAbvGr, BldgType,
## BsmtExposure, BsmtFinSF1, BsmtFinType1 and 44 more;
##  27 attributes confirmed unimportant: Artery, BsmtFinSF2,
## BsmtHalfBath, ExterCond, Feedr and 22 more;
##  10 tentative attributes left: Alley, BsmtCond, BsmtFinType2,
## EnclosedPorch, Fence and 5 more;
```

```
plot(fit.bt)
```




## Summary

The plot below shows the (centered, scaled) importance ranking for features not rejected by Boruta. I am not sure if we should perform model tuning/selection on all of these features, or choose a subset to investigate interactions or higher order terms. I also did not center and scale the design matrixes here.

```
# Look at variables not rejected by Boruta.
keep <- names(fit.bt$finalDecision[fit.bt$finalDecision != 'Rejected'])
length(keep)
```

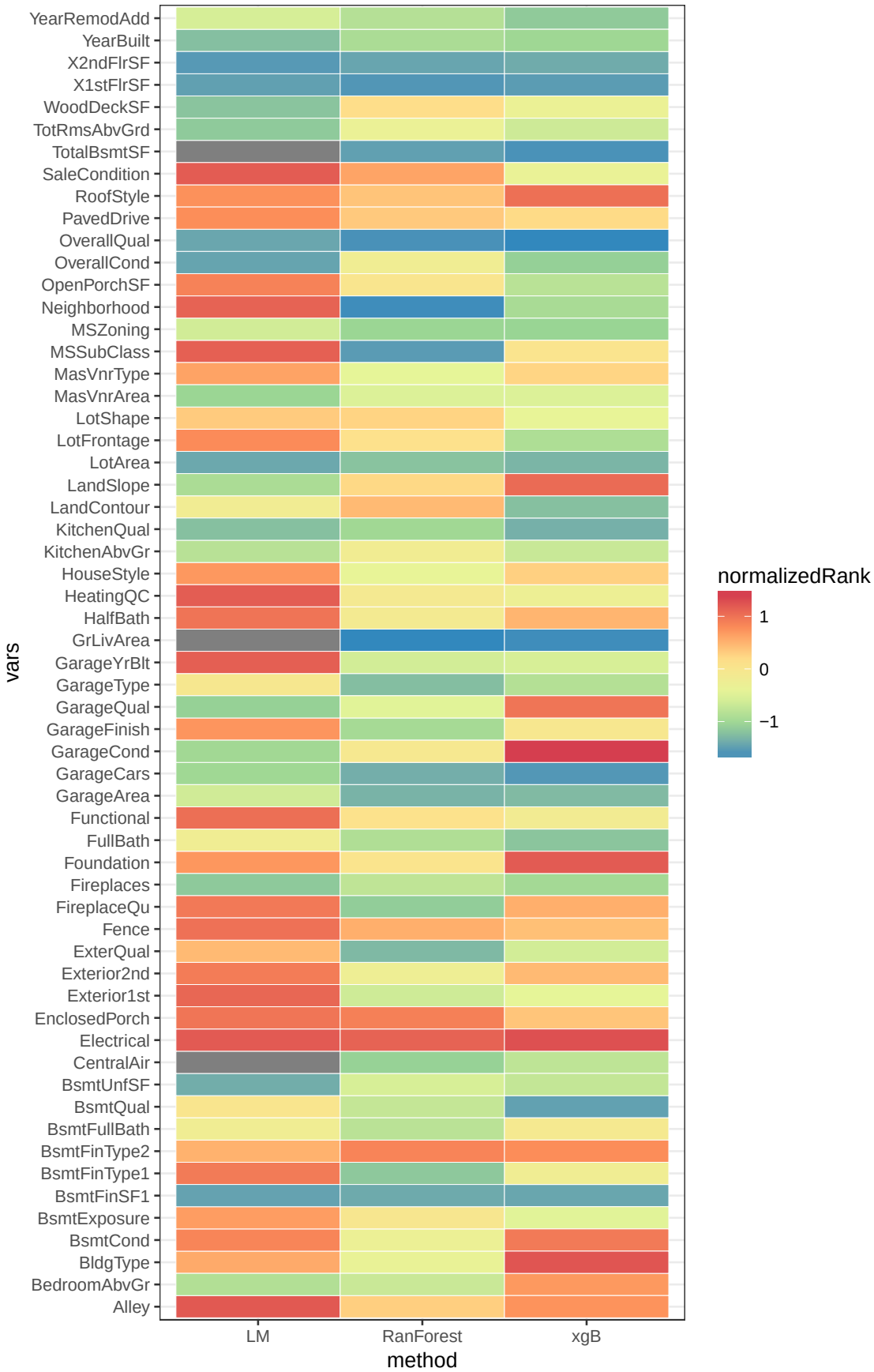
```
## [1] 59
```

```
# Deal with factor indicators -- take max rank over factor levels.
lmranks <- rank(-imp.lm$Overall)
lmnames <- rownames(imp.lm)
maxrank <- sapply(keep, function(i) {
```

```
  hits <- sapply(lmnames, function(j) grepl(i, j))
  if (sum(hits) == 0)
    return(NA)
  if (sum(hits) == 1) {
    sel <- (lmnames == i)
    if (sum(sel) == 0)
      return(NA)
    return(lmranks[sel])
  }
  levs <- lmnames[hits]
  idx <- lmnames %in% levs
  res <- max(lmranks[idx])
  return(res)
})

ranks <- c(scale(unlist(maxrank)),
           scale(rank(-imp.rf$Overall)),
           scale(rank(-imp.bst$Gain)))
vars <- c(names(maxrank), rownames(imp.rf), imp.bst$Feature)
method <- rep(c('LM', 'RanForest', 'xgB'),  c(length(maxrank), nrow(imp.rf), nrow(imp.bst)))
pd <- data.frame(normalizedRank = ranks, vars, method)
```

```
ggplot(subset(pd, vars %in% keep), aes(method, vars)) +
    geom_tile(aes(fill = normalizedRank), colour = "white") +
        scale_fill_distiller(palette = 'Spectral') +
            theme_bw()
```


YearRemodAdd
YearBuilt
X2ndFlrSF
X1stFlrSF
WoodDeckSF
TotRmsAbvGrd
TotalBsmtSF
SaleCondition
RoofStyle
PavedDrive
OverallQual
OverallCond
OpenPorchSF
Neighborhood
MSZoning
MSSubClass
MasVnrType
MasVnrArea
LotShape
LotFrontage
LotArea
LandSlope
LandContour
KitchenQual
KitchenAbvGr
HouseStyle
HeatingQC
HalfBath
GrLivArea
GarageYrBlt
GarageType
GarageQual
GarageFinish
GarageCond
GarageCars
GarageArea
Functional
FullBath
Foundation
Fireplaces
FireplaceQu
Fence
ExterQual
Exterior2nd
Exterior1st
EnclosedPorch
Electrical
CentralAir
BsmtUnfSF
BsmtQual
BsmtFullBath
BsmtFinType2
BsmtFinType1
BsmtFinSF1
BsmtExposure
BsmtCond
BldgType
BedroomAbvGr
Alley
vars
LM
RanForest
xgB
method
normalizedRank
1
0
−1