Homework 1

Yufei Lin, Jingfeng Xia

Nov 7 2020

```
## [1] "4"
```

Variable Name: 1. HousePricing:1400 dataset 2. train: training data

Data Processing

Read in Data

We have chosen to eliminate the ID column from this dataset because ID has nothing to do with our prediction and would mess up our prediction.

Pairs of Categories (Iris)

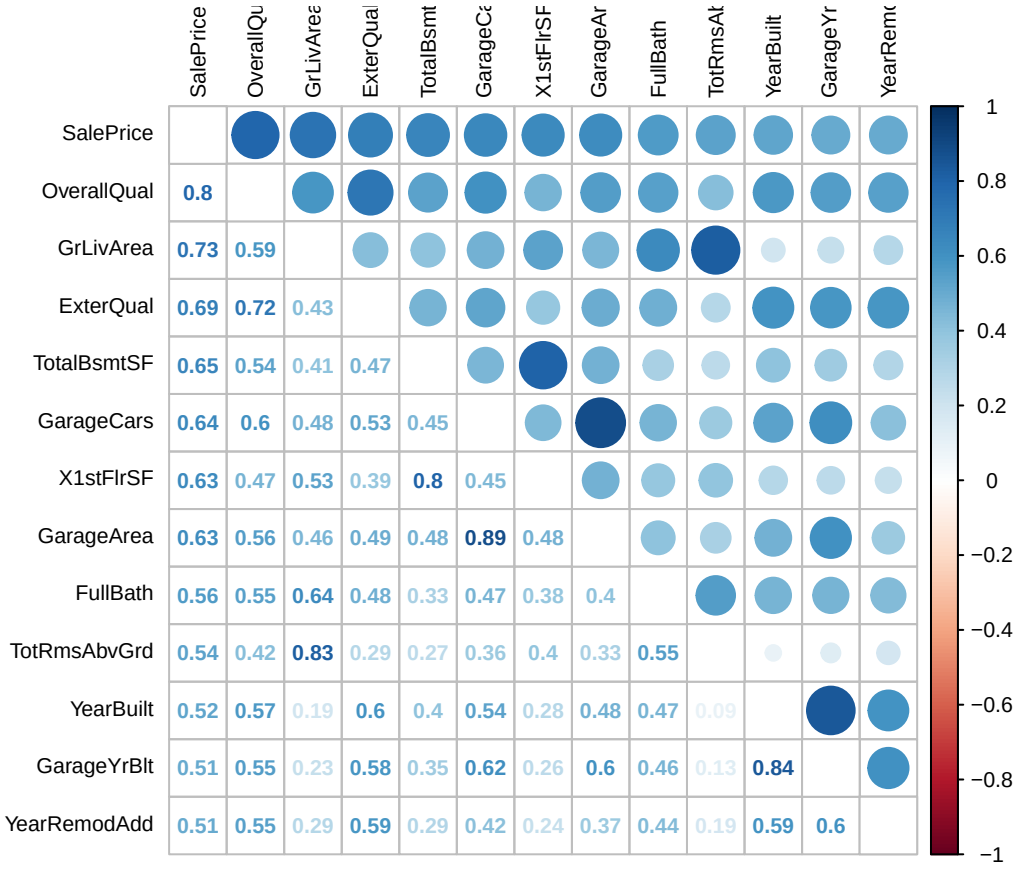
```
## [1] "Hi"
```

Feature Engineering

In this section, we convert all missing value based on the following rules:

1. Categorical: fill in most common
2. Numeric: fill in median/average

Convert all train to HousePricing



Boostraping

Seperate into Test and Training Set

Spearate by 70% train, 30% test.

Convert to CSV upload

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  131000  168000  182946  219210  745000
```

```
##   0   1
## 538 483
```

Hierachical

Each team member bootstraps training data.

Prediction Algorithms

Each model needs a cross validation algorithm Remember to report RMSE

Regression Methods

1. PCR (Iris)

Cross Validation

2. Random Forest (Yufei Lin)

Prepare Model

We have 199 independent variables in the data set, therefore we have set mtry(Number of randomly selected variables for each split) to be the square root of that number for maximum performance of the model.

Call: randomForest(formula = SalePrice ~ ., data = train, mtry = sqrt(totalIV) + 1000, importance = TRUE) Type of random forest: regression Number of trees: 500 No. of variables tried at each split: 52

Mean of squared residuals: 483805057 % Var explained: 92.23

Check Accuracy

If we raise the allowance to $$$10000 We could raise the prediction accuracy to almost 95% of a time.

```
##        1        2        3        4        5        6
## 128866.2 247184.2 258383.7 150744.4 223634.8 286869.7
```

```
## [1]  99500 215000 275000 143000 239000 290000
```

```
## [1] 0.5535308
```

Error Metrics

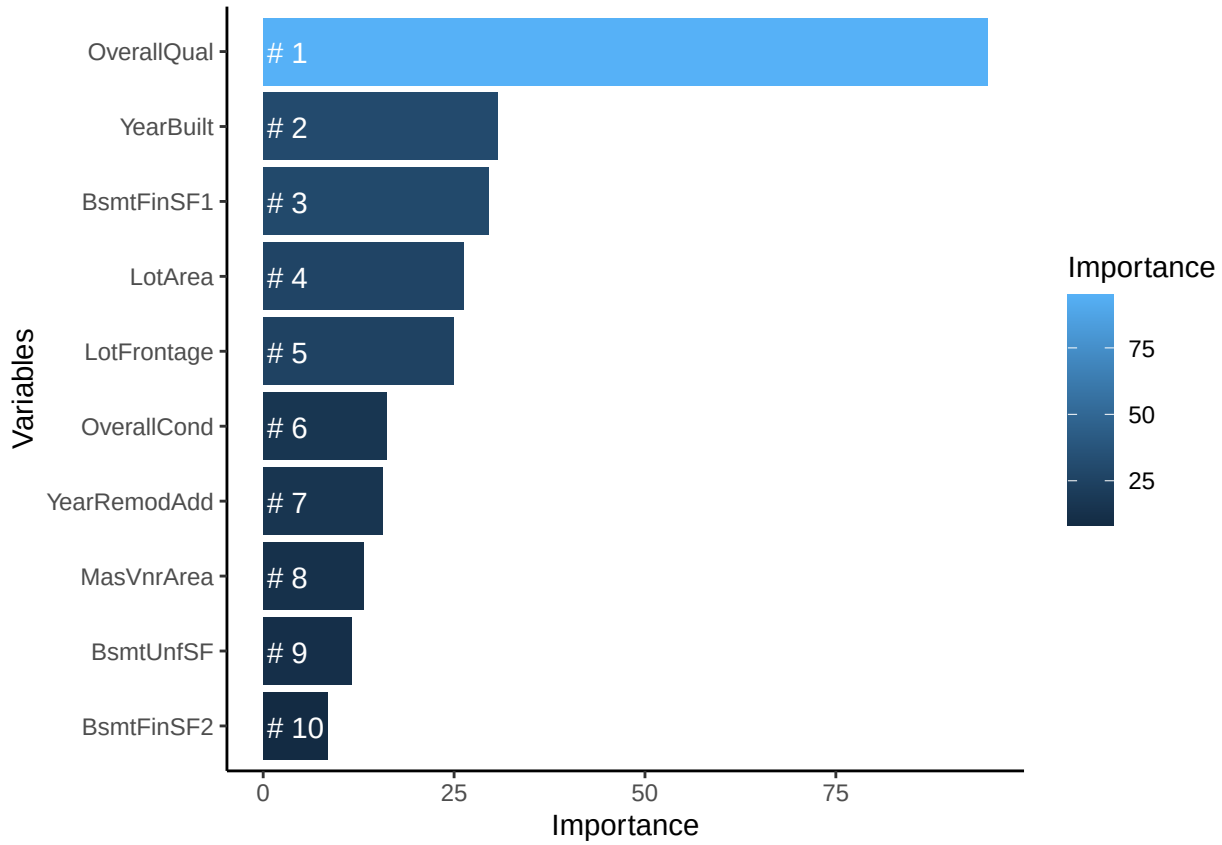
Then let us take a look at the MSE of this model:

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

```
## [1] 738057016
```

Variable Importance

Here we are going to show the top 10 most important variables in predicting sale price of a house.



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

1. OverallQual
2. LotArea
3. GrLivArea
4. Neighbourhood

Cross Validation

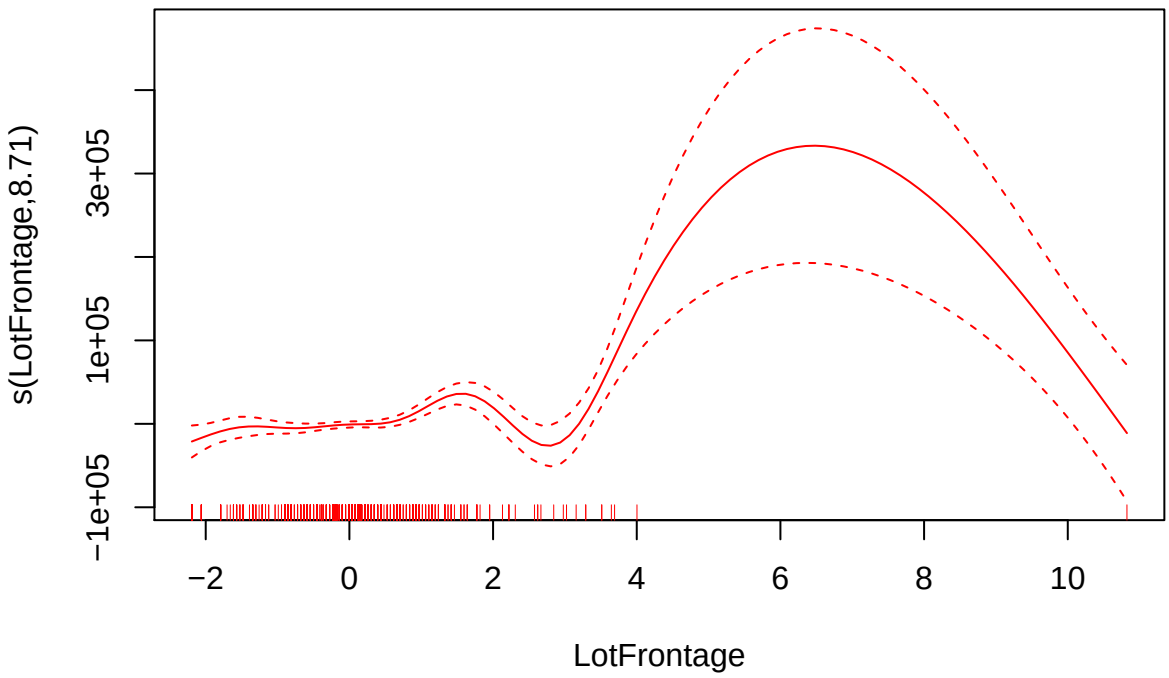
```
##          R2     RMSE      MAE
## 1 0.8627518 27167.2 12341.62
```

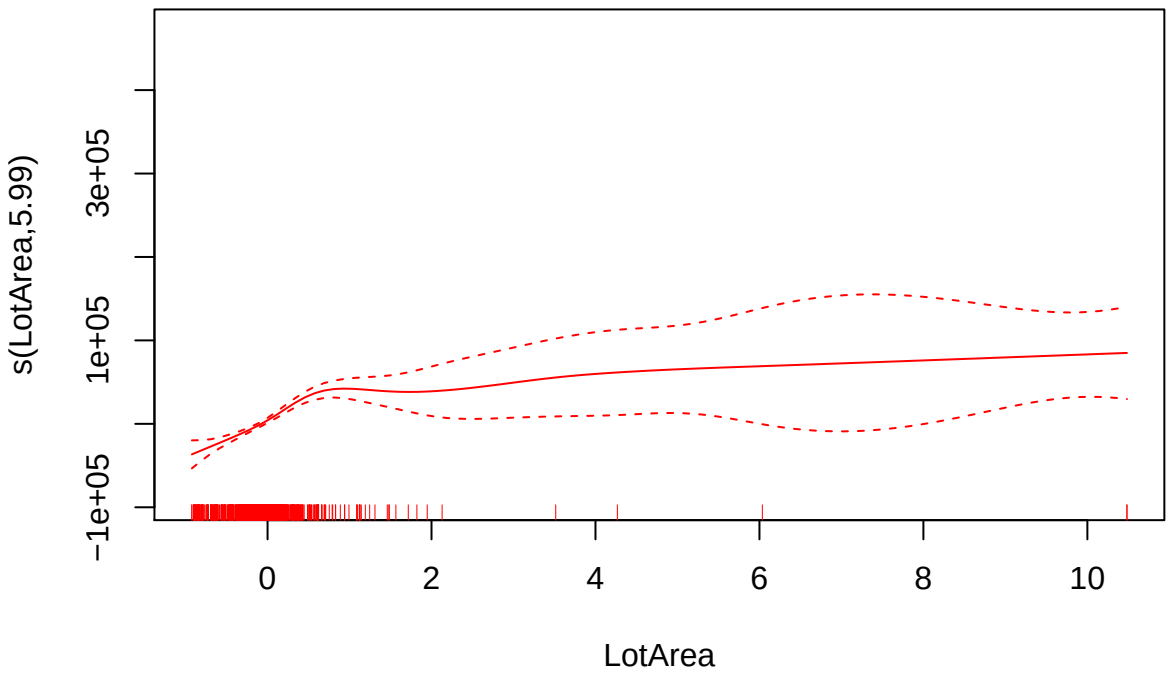
Therefore, we will make GAM models according to these three factors.

3. GAM (Yufei Lin)

1) GAM1

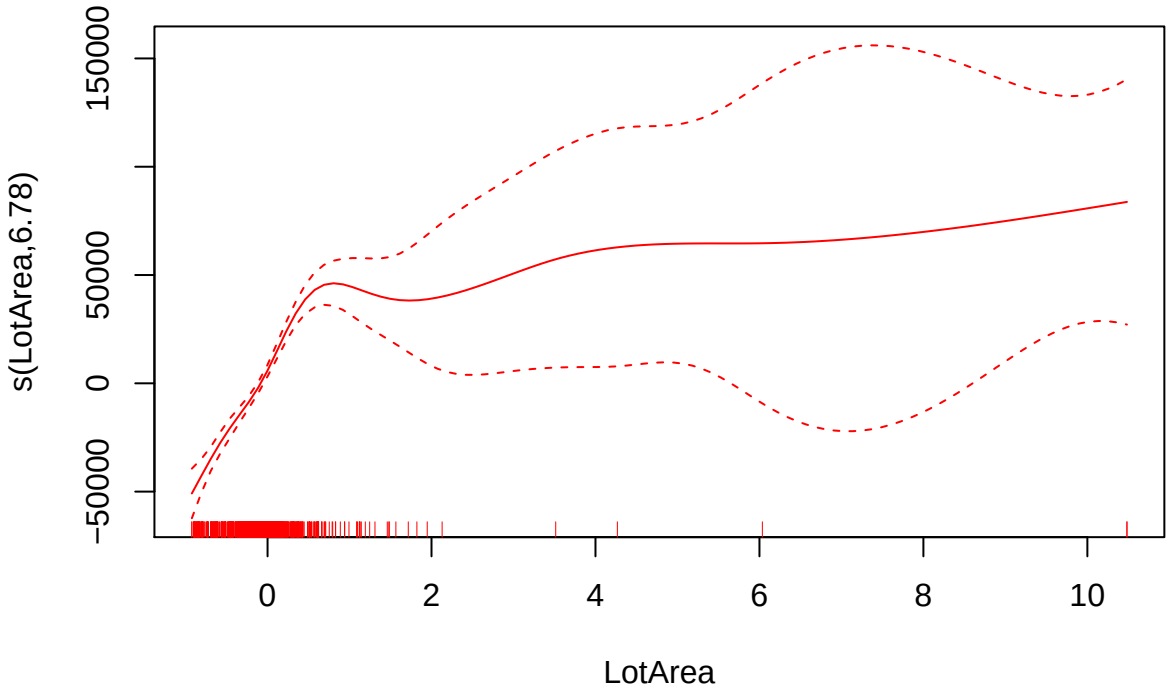
```
## [1] "Deviance of Model 1"
## [1] 1.53569e+12
## [1] "Accuracy of Model 1"
## [1] 0.1936219
```


s(LotFrontage,8.71)
−1e+05
1e+05
3e+05
−2
0
2
4
6
8
10
LotFrontage



2) GAM1

```
## [1] "Deviance of Model 2"
## [1] 1.628086e+12
## [1] "Accuracy of Model 2"
## [1] 0.1776765
```



3) GAM3

```
## [1] "Deviance of Model 3"
```

```
## [1] 1.779948e+12
```

```
## [1] "Accuracy of Model 3"
```

```
## [1] 0.2118451
```

GAM Summary

```
## Analysis of Deviance Table
##
## Model 1: SalePrice ~ s(LotFrontage) + s(LotArea) + OverallQual
## Model 2: SalePrice ~ OverallCond + s(LotArea) + OverallQual
## Model 3: SalePrice ~ LotFrontage + YearRemodAdd + OverallQual
##   Resid. Df Resid. Dev      Df    Deviance       F    Pr(>F)
## 1    1002.9 1.5357e+12
## 2    1010.2 1.6281e+12 -7.2419 -9.2396e+10  8.3437 3.416e-10 ***
## 3    1017.0 1.7799e+12 -6.8385 -1.5186e+11 14.5226 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

Cross Validation

```
## [1] "Cross Validation of Model 1"
```

```
##          R2     RMSE      MAE
## 1 0.7038757 40360.77 29481.19
```

```
## [1] "Cross Validation of Model 2"
```

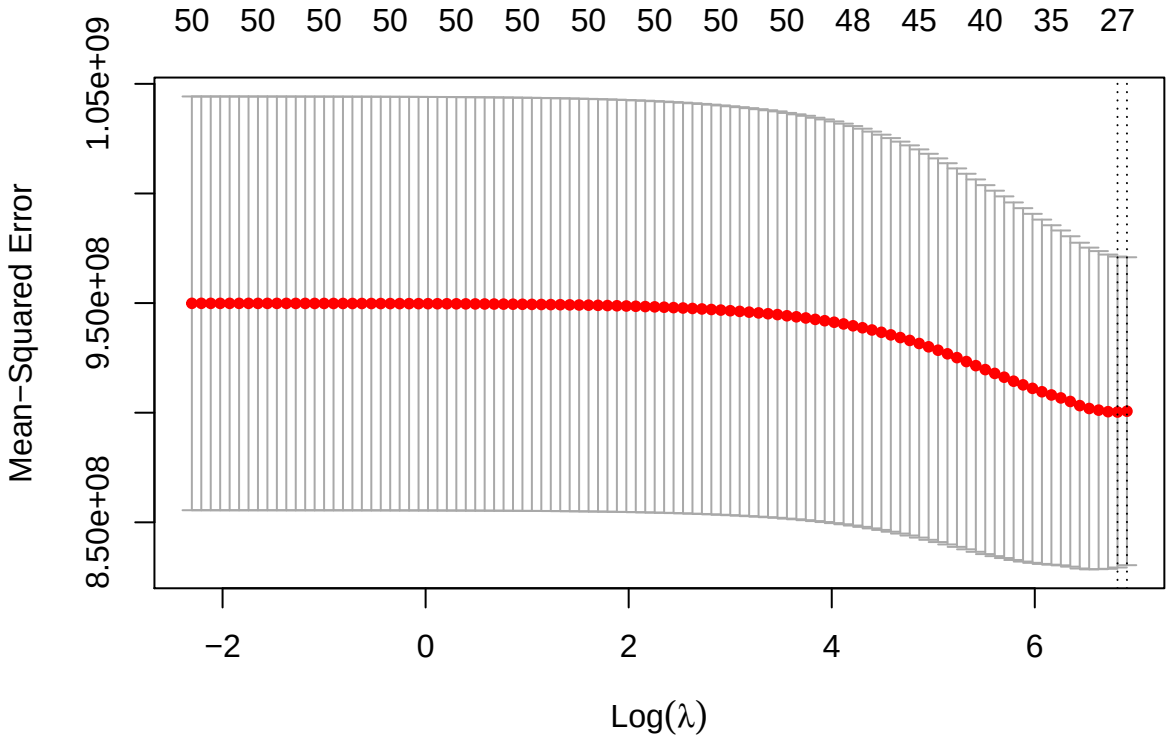
```
##          R2     RMSE      MAE
## 1 0.7120391 39424.05 29120.45
```

```
## [1] "Cross Validation of Model 3"
```

```
##          R2     RMSE     MAE
## 1 0.6722334 42156.84 30208.6
```

4. Lasso & Ridge (Jinhong)

```
## [1] 911.1628
```




```
## [1] 33310.11
```

```
## 53 x 1 sparse Matrix of class "dgCMatrix"
##                            1
## (Intercept)   180624.4714
## LotFrontage     3121.9309
## LotArea         4072.8449
## OverallQual    24975.5632
## OverallCond     3276.8118
## YearBuilt       6187.7115
## YearRemodAdd    4844.0623
## MasVnrArea      2340.8189
```

```
## BsmtFinSF1      9260.1149
## BsmtFinSF2         .
## BsmtUnfSF          .
## TotalBsmtSF    10974.3922
## X1stFlrSF          .
## X2ndFlrSF          .
## LowQualFinSF    -555.3530
## GrLivArea      31811.6217
## BsmtFullBath       .
## BsmtHalfBath       .
## FullBath           .
## HalfBath           .
## BedroomAbvGr   -7371.7812
## KitchenAbvGr   -2231.7620
## TotRmsAbvGrd     867.3237
## Fireplaces       976.0580
## GarageYrBlt    -3146.4087
## GarageCars      4103.5761
## GarageArea      5123.5766
## WoodDeckSF         .
## OpenPorchSF        .
## EnclosedPorch      .
## X3SsnPorch         .
## ScreenPorch        .
## PoolArea        1553.9794
## MiscVal            .
## MoSold             .
## YrSold             .
## MSSubClass20       .
## MSSubClass30     465.6329
## MSSubClass40       .
## MSSubClass45       .
## MSSubClass50     775.8989
## MSSubClass60       .
## MSSubClass70       .
## MSSubClass75  -10379.7336
## MSSubClass80       .
## MSSubClass85       .
## MSSubClass90   -1345.0041
## MSSubClass120  -7051.8454
## MSSubClass160 -17684.2305
## MSSubClass180      .
## MSSubClass190      .
## AlleyNone       2625.6437
## AlleyPave      -6504.7948
```

Cross Validation

5. Splines (Jingfeng)

Cross Validation

6. Linear Regression (Yanze)

```
##
## Call:
## lm(formula = SalePrice ~ ., data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -99485 -15981  -2018  14134 199147
##
## Coefficients: (3 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   183432.84   10130.59  18.107  < 2e-16 ***
## LotFrontage     3147.84    1289.54   2.441  0.01482 *
## LotArea         4836.74    1573.41   3.074  0.00217 **
## OverallQual    23047.99    1715.50  13.435  < 2e-16 ***
## OverallCond     5470.20    1165.09   4.695 3.05e-06 ***
## YearBuilt      15518.24    2486.13   6.242 6.45e-10 ***
## YearRemodAdd    3137.86    1422.37   2.206  0.02761 *
## MasVnrArea      2908.01    1068.12   2.723  0.00659 **
## BsmtFinSF1     21248.18    2322.14   9.150  < 2e-16 ***
## BsmtFinSF2      5026.86    1108.59   4.534 6.50e-06 ***
## BsmtUnfSF      10614.00    1938.86   5.474 5.59e-08 ***
## TotalBsmtSF           NA         NA      NA       NA
## X1stFlrSF      22404.89    2456.62   9.120  < 2e-16 ***
## X2ndFlrSF      31466.67    2832.58  11.109  < 2e-16 ***
## LowQualFinSF    1380.58    1149.28   1.201  0.22995
## GrLivArea             NA         NA      NA       NA
## BsmtFullBath    -333.09    1415.66  -0.235  0.81403
## BsmtHalfBath     289.86    1019.07   0.284  0.77614
## FullBath        -983.22    1551.75  -0.634  0.52648
## HalfBath        -346.58    1378.32  -0.251  0.80152
## BedroomAbvGr  -10464.45    1446.51  -7.234 9.49e-13 ***
## KitchenAbvGr   -1931.61    2078.57  -0.929  0.35297
## TotRmsAbvGrd    4734.30    2067.28   2.290  0.02223 *
## Fireplaces      1221.48    1207.05   1.012  0.31181
## GarageYrBlt    -5320.93    1235.08  -4.308 1.81e-05 ***
## GarageCars      6002.22    2325.53   2.581  0.01000 **
## GarageArea      4287.50    2212.87   1.938  0.05297 .
## WoodDeckSF       185.87    1107.84   0.168  0.86680
## OpenPorchSF      994.30    1019.96   0.975  0.32988
## EnclosedPorch    643.97    1089.62   0.591  0.55466
## X3SsnPorch        60.99     830.84   0.073  0.94150
## ScreenPorch      720.30     907.27   0.794  0.42744
## PoolArea        2221.55     945.34   2.350  0.01897 *
## MiscVal         -547.47     864.79  -0.633  0.52684
## MoSold           708.67     896.80   0.790  0.42959
## YrSold          -597.24     928.95  -0.643  0.52043
## MSSubClass20     195.23    9136.38   0.021  0.98296
## MSSubClass30   14253.24    9565.77   1.490  0.13654
## MSSubClass40    4436.02   16793.99   0.264  0.79173
## MSSubClass45   10119.91   12288.69   0.824  0.41042
## MSSubClass50    5803.53    8976.91   0.646  0.51811
## MSSubClass60  -16743.13   10020.15  -1.671  0.09505 .
```

```
## MSSubClass70    8319.08    9988.10   0.833  0.40511
## MSSubClass75  -14707.99   13659.15  -1.077  0.28184
## MSSubClass80   -2800.34   10188.16  -0.275  0.78348
## MSSubClass85    1322.14   11964.88   0.111  0.91203
## MSSubClass90   -8268.50   10204.63  -0.810  0.41798
## MSSubClass120 -15412.36   10099.31  -1.526  0.12732
## MSSubClass160 -30393.38   10410.93  -2.919  0.00359 **
## MSSubClass180  -6075.76   17114.17  -0.355  0.72266
## MSSubClass190        NA         NA      NA       NA
## AlleyNone       3079.38    5977.20   0.515  0.60654
## AlleyPave      -7357.74    7759.07  -0.948  0.34322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28040 on 971 degrees of freedom
## Multiple R-squared:   0.88,  Adjusted R-squared:  0.8739
## F-statistic: 145.3 on 49 and 971 DF,  p-value: < 2.2e-16
```

```
##          1          2          3          4          5          6          7
## 104538.668 237272.277 258836.311 186095.762 281258.115 295303.393 179935.000
##          8          9         10         11         12         13         14
## 139986.803 274613.235 150848.750 301854.028 143930.805 221986.778 125299.144
##         15         16         17         18         19         20         21
##  75336.342 201627.997 140562.916 162289.838 129102.814 294457.596 195681.483
##         22         23         24         25         26         27         28
##  82700.998 238883.526 239164.835  95779.513 161067.198  72559.045 210805.829
##         29         30         31         32         33         34         35
## 210729.373  87236.813 141565.127 212838.659  57906.745 111697.974 405143.418
##         36         37         38         39         40         41         42
## 186345.099 285091.663 183464.766 172318.939 101358.206 123415.583 122855.964
##         43         44         45         46         47         48         49
##  58994.263 158516.325   4296.359 212009.098 140838.864 180230.895 168739.077
##         50         51         52         53         54         55         56
## 165442.257 248101.739 328579.613 106176.824 217592.762 301722.960 316582.003
##         57         58         59         60         61         62         63
## 207502.950 344053.614 195385.285 174647.042 195938.380 172079.707 102471.778
##         64         65         66         67         68         69         70
## 198662.378 186345.099 196770.469 102280.262 198202.134 180783.955 251797.719
##         71         72         73         74         75         76         77
## 275070.757 103212.370  85482.692 200958.700 130422.129 180700.002 208323.574
##         78         79         80         81         82         83         84
## 207114.525 133429.451 196275.446 276956.269 176672.259 109881.632 158010.788
##         85         86         87         88         89         90         91
## 144581.963 158516.325 196893.064 174269.761 161860.820 279645.815 134451.502
##         92         93         94         95         96         97         98
## 123507.060 181212.075 289065.800 181871.109 288899.293 147305.180 183291.985
##         99        100        101        102        103        104        105
##  52126.885  99288.112 185156.050 148016.839 149920.668 140525.467 145498.879
##        106        107        108        109        110        111        112
## 117341.767 210729.373 214243.954 189076.611 295255.748 259737.009 173029.891
##        113        114        115        116        117        118        119
## 106176.824 169380.987 132695.192 103576.610 140565.080 221979.800 326574.876
##        120        121        122        123        124        125        126
## 158883.836 182140.371 154725.130 192345.317 379742.456 126872.224 271879.608
```

```
##        127        128        129        130        131        132        133
## 238883.526 181798.346 328579.613 135628.199 178971.183 253483.778 206377.336
##        134        135        136        137        138        139        140
## 130201.941  83166.273  66365.294 123470.813 148668.427 164529.193 252119.743
##        141        142        143        144        145        146        147
## 126756.249 239891.545 160641.005 246783.553 177777.847 200982.595 141490.052
##        148        149        150        151        152        153        154
## 105779.007 194901.633 235743.262 259895.218 134102.401  75336.342 297263.057
##        155        156        157        158        159        160        161
## 111067.823 159891.170  93996.787  92699.077  72559.045 479951.971 128496.558
##        162        163        164        165        166        167        168
## 246832.589 184125.871 235256.822 199987.985  67163.650 148246.649 125522.005
##        169        170        171        172        173        174        175
## 204679.785 204048.013 147699.718 255416.265 211064.417 120438.057 227304.754
##        176        177        178        179        180        181        182
## 122855.964  21755.691  68785.846 113825.528 281095.114 127254.316 124012.315
##        183        184        185        186        187        188        189
## 184992.841 609887.201 312844.167  92699.077 119065.985 223969.228 137502.552
##        190        191        192        193        194        195        196
## 149719.075 321410.098 144039.614 243134.104 270539.260 192518.497 163705.579
##        197        198        199        200        201        202        203
## 316777.714 221851.329 142764.421 288771.601  57837.212 127666.337 136359.752
##        204        205        206        207        208        209        210
## 130422.129  64678.459  92379.618  93855.279 154735.566 205461.509 445407.161
##        211        212        213        214        215        216        217
## 276997.618 151488.810 132931.710 120514.366 153190.660 183464.766 159846.470
##        218        219        220        221        222        223        224
## 277009.794 144019.705 182565.807  88590.030 261043.895 130368.430  92379.762
##        225        226        227        228        229        230        231
## 371891.355 239512.012  82264.582 157494.421 271845.994 128638.439 199987.985
##        232        233        234        235        236        237        238
## 316777.714  91101.983 344284.021 187619.922 199225.612 176973.121 389047.718
##        239        240        241        242        243        244        245
##  73868.443 263679.475 242228.577 147947.372 220467.372 129102.814 109881.632
##        246        247        248        249        250        251        252
## 357793.331 144578.028 129102.814 175012.605 143443.888 143586.595  99303.771
##        253        254        255        256        257        258        259
## 338396.503 118263.588 378210.780 131604.242 312555.325 170580.378 144039.614
##        260        261        262        263        264        265        266
##  96229.788 178447.626 102280.262 259895.218  84981.856 374623.926 221851.329
##        267        268        269        270        271        272        273
##  20048.462 137479.838 109276.519 166347.809 134225.828 157494.421 259457.539
##        274        275        276        277        278        279        280
## 189076.611 122855.964 120255.560 111706.663 318324.335 150507.526 135993.510
##        281        282        283        284        285        286        287
## 173264.052 103616.606 209074.118 117280.905 155882.950 159558.680 181871.109
##        288        289        290        291        292        293        294
## 122189.707 172318.939 289053.006 207114.525  52974.981 147699.718 208389.914
##        295        296        297        298        299        300        301
## 225891.681 192809.514 186095.762 105153.288 342790.562 228484.923 130503.613
##        302        303        304        305        306        307        308
## 187038.963 305865.536 209995.511 103480.719 154725.130 207579.715 172468.733
##        309        310        311        312        313        314        315
## 251797.719 186725.042 150000.965  66365.294 204051.645 165754.171 130031.538
```

```
##        316        317        318        319        320        321        322
## 210136.808 204324.610 131812.618 207544.419 105330.561 178039.966 188258.198
##        323        324        325        326        327        328        329
## 201229.259 137963.095  21755.691 301382.159 237885.466 118550.708 213718.854
##        330        331        332        333        334        335        336
## 146709.913 273265.750  73673.837 132931.710 172713.530 106094.081 119608.369
##        337        338        339        340        341        342        343
##  91814.769 187038.963 111098.993 164528.664  70225.641 145515.978 164856.351
##        344        345        346        347        348        349        350
## 160678.388 285091.663 126081.209 256891.489 160678.388 235818.972 189807.612
##        351        352        353        354        355        356        357
## 118401.391 120871.777 125731.196 225275.542 146505.849 124693.161 181212.075
##        358        359        360        361        362        363        364
## 200741.564 226218.564 130422.129  67163.650  29406.117 204475.412 157637.254
##        365        366        367        368        369        370        371
## 103576.610 254563.937  71265.302 128462.667 202318.231 261005.683 213159.427
##        372        373        374        375        376        377        378
## 187977.773 111706.663 171531.108 160968.103 107560.361 123532.834 213201.209
##        379        380        381        382        383        384        385
## 215700.360 178369.564 180783.955 234197.462 326565.449 131452.067 133993.592
##        386        387        388        389        390        391        392
## 121068.608 125585.156 215834.890 156964.870 282467.613 210713.115 159769.340
##        393        394        395        396        397        398        399
## 138226.664 123080.335 235208.613  87894.484 221346.128 131812.618 229569.351
##        400        401        402        403        404        405        406
## 204048.013  99674.316 305865.536 156005.545 230712.058 188527.648 100443.926
##        407        408        409        410        411        412        413
##  92379.618 214295.357 187121.130 259457.539 151044.524 319939.870 125046.822
##        414        415        416        417        418        419        420
## 210614.485 197303.051 166750.133 255416.265 138447.581  96482.979 164528.664
##        421        422        423        424        425        426        427
## 236070.296 158832.110 271879.608 131026.540 127733.048 177558.057  93855.279
##        428        429        430        431        432        433        434
##  77786.121 222023.277 159314.012 321457.445 260520.097 165414.542 184996.884
##        435        436        437        438        439
## 217592.762 183291.985  46412.783 154321.459 144003.089
```

Cross Validation

7. Ensemble

Evaluation of different models

Root MSE

Choose best fit model

Discussion & Future Development

Resources

https://www.kaggle.com/erikbruin/house-prices-lasso-xgboost-and-a-detailed-eda