Comparing Models for the Prediction of Home Prices

Eric Hirsch, Carlisle Ferguson, and Cameron Smith

5/15/2022

Contents

Abstract

Being able to accurately predict housing prices is critical to many industries. Recently, analysts have attempted to improve price prediction with enhanced statistical techniques. In this paper, we take a more comparative approach, examining 7 regression techniques (OLS, ridge, lasso, elastic net, simple decision tree, random forest and gradient boosting) to assess the best performance. We used a kaggle dataset (https://www.kaggle.com/c/house-prices-advanced-regression-techniques) in order to test the performance of the model. We found gradient boosting to be the best predictor, as is often the case because of the machine-learning algorithm at the heart of gradient boosting. Lasso was the best non-tree predictor, which we speculate is because the dataset has a high number of predictors relative to the number of observations.

Keywords: Regression, OLS, Ridge, Lasso, elastic net, random forest, gradient boosting, home prices

Introduction

In this paper we analyze housing prices by comparing various prediction methodologies: OLS, ridge, lasso, elastic net, simple decision tree, random forest and gradient boosting. The purpose is to compare the methodologies and draw conclusions about which are most effective and why. Regression alone is not necessarily the optimal strategy for predicting housing prices.[1] However, when data sets and/or analysis resources are limited, regression can perform adequately.

Background and Literature Review

The ability to accurately predict home prices is of tremendous value to a number of industries, including investors, real estate agents, and municipalities who depend upon property tax revenue. Predictive models

[1] 1 Li, 2021

for home prices fall roughly into two kinds. First, there are those which predict market trends, busts, and booms. These predictions rely mainly on time series data and analysis of housing prices in the aggregate. The other type of prediction involves the capacity to predict individual house prices from a set of factors. These usually employ some form of regression and/or machine learning.[2]

For either sort of prediction, there is no consensus about the best method. Many researchers have sought to enhance the traditional models with other methodologies.[3] For example, Guan et. al. propose a "data stream" approach in which past sale records are treated as an evolving datastream.[4] Li et. al. introduce a "grey seasonal model" in which seasonal fluctuations are modeled using grey systems theory, which incorporates uncertainty.[5] Alfiyatin, et. el. use particle swarm optimization (PSO) to select independent variables.[6] (PSO is an optimization system in which population is initialized with random solutions and searches for optima by updating generations.) Finally, Liu et.al incorporate both spatial and temporal autocorrelation in their models by analyzing experience-based submarkets identified by real estate professionals.[7]

All of these researchers report that their innovations improve their regression models. Indeed, any real estate agent can tell you that a predictive model can be improved simply by knowing what other houses in the neighborhood sold for. The problem is, the data at the center of these enhancements is not always available. The researcher may have home sales from only a short time span, and neighborhoods that are not defined by real estate experts but by traditional boundary lines which may contain a mix of house types. Even when data is available, the complex models proposed may be computationally expensive and/or require data analysis expertise that is not generally available.

In this project we approach the question comparatively. Restricting ourselves to regression models, we compare seven types of regression: OLS, ridge, lasso, elastic net, decision tree, random forest and gradient boosting. The data is drawn from the Advanced Regression Techniques housing data set for Ames, Iowa. We test the accuracy of our models by submitting each to the Kaggle competition to see how they perform. We then discuss the merits of the different sorts of approaches.

Modeling

We are modeling a data set containing 1460 records of houses sold in the Ames, Iowa area between 2006 and 2010. The variables are mostly related to house features, such as square footage, the presence of a pool,

[2] 2 Journal, 2019
[3] 3 Wu, 2020
[4] 4 Guan, 2021
[5] 5 Li, 2021
[6] 6 Alfiyatin, 2017
[7] 7 Liu, X. 2012

etc. The response variable, "SalePrice", is a continuous variable representing the sale price of the house in thousands of dollars.

We examine the data:

1. Dataset Description

A. Summary Statistics

```
##        Id           MSSubClass       MSZoning     LotFrontage
##  Min.   :   1.0   Min.   : 20.0   C (all):  10   Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   FV     :  65   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   RH     :  16   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9   RL     :1151   Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0   RM     : 218   3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                   Max.   :313.00
##                                                  NA's   :259
##     LotArea        Street      Alley      LotShape  LandContour  Utilities
##  Min.   :  1300   Grvl:   6   Grvl:  50   IR1:484   Bnk:  63    AllPub:1459
##  1st Qu.:  7554   Pave:1454   Pave:  41   IR2: 41   HLS:  50    NoSeWa:   1
##  Median :  9478               NA's:1369   IR3: 10   Low:  36
##  Mean   : 10517                           Reg:925   Lvl:1311
##  3rd Qu.: 11602
##  Max.   :215245
##
##    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      YearBuilt
```

```
##  1Fam   :1220   1Story :726   Min.   : 1.000   Min.   :1.000   Min.   :1872
##  2fmCon:  31   2Story :445   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954
##  Duplex:  52   1.5Fin :154   Median : 6.000   Median :5.000   Median :1973
##  Twnhs :  43   SLvl   : 65   Mean   : 6.099   Mean   :5.575   Mean   :1971
##  TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000
##                1.5Unf : 14   Max.   :10.000   Max.   :9.000   Max.   :2010
##                (Other): 19
##   YearRemodAdd    RoofStyle       RoofMatl     Exterior1st   Exterior2nd
##  Min.   :1950   Flat   :  13   CompShg:1434   VinylSd:515   VinylSd:504
##  1st Qu.:1967   Gable  :1141   Tar&Grv:  11   HdBoard:222   MetalSd:214
##  Median :1994   Gambrel:  11   WdShngl:   6   MetalSd:220   HdBoard:207
##  Mean   :1985   Hip    : 286   WdShake:   5   Wd Sdng:206   Wd Sdng:197
##  3rd Qu.:2004   Mansard:   7   ClyTile:   1   Plywood:108   Plywood:142
##  Max.   :2010   Shed   :   2   Membran:   1   CemntBd: 61   CmentBd: 60
##                                (Other):   2   (Other):128   (Other):136
##    MasVnrType    MasVnrArea     ExterQual ExterCond  Foundation  BsmtQual
##  BrkCmn : 15   Min.   :   0.0   Ex: 52    Ex:   3   BrkTil:146   Ex  :121
##  BrkFace:445   1st Qu.:   0.0   Fa: 14    Fa:  28   CBlock:634   Fa  : 35
##  None   :864   Median :   0.0   Gd:488    Gd: 146   PConc :647   Gd  :618
##  Stone  :128   Mean   : 103.7   TA:906    Po:   1   Slab  : 24   TA  :649
##  NA's   :  8   3rd Qu.: 166.0             TA:1282   Stone :  6   NA's: 37
##                Max.   :1600.0                       Wood  :  3
##                NA's   :8
##  BsmtCond    BsmtExposure BsmtFinType1   BsmtFinSF1     BsmtFinType2
##  Fa  :  45   Av  :221     ALQ :220     Min.   :   0.0   ALQ :  19
##  Gd  :  65   Gd  :134     BLQ :148     1st Qu.:   0.0   BLQ :  33
##  Po  :   2   Mn  :114     GLQ :418     Median : 383.5   GLQ :  14
##  TA  :1311   No  :953     LwQ : 74     Mean   : 443.6   LwQ :  46
##  NA's:  37   NA's: 38     Rec :133     3rd Qu.: 712.2   Rec :  54
##                           Unf :430     Max.   :5644.0   Unf :1256
##                           NA's: 37                      NA's:  38
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741
```

```
##  1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49
##  Median :   0.00   Median : 477.5   Median : 991.5   GasW :  18   Gd:241
##  Mean   :  46.55   Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
##  Max.   :1474.00   Max.   :2336.0   Max.   :6110.0   Wall :   4
##
##  CentralAir Electrical     X1stFlrSF      X2ndFlrSF     LowQualFinSF
##  N:  95     FuseA:  94   Min.   : 334   Min.   :   0   Min.   :  0.000
##  Y:1365     FuseF:  27   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000
##             FuseP:   3   Median :1087   Median :   0   Median :  0.000
##             Mix  :   1   Mean   :1163   Mean   : 347   Mean   :  5.845
##             SBrkr:1334   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000
##             NA's :   1   Max.   :4692   Max.   :2065   Max.   :572.000
##
##    GrLivArea     BsmtFullBath     BsmtHalfBath        FullBath
##  Min.   : 334   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:1130   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000
##  Median :1464   Median :0.0000   Median :0.00000   Median :2.000
##  Mean   :1515   Mean   :0.4253   Mean   :0.05753   Mean   :1.565
##  3rd Qu.:1777   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000
##  Max.   :5642   Max.   :3.0000   Max.   :2.00000   Max.   :3.000
##
##     HalfBath       BedroomAbvGr    KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100      Min.   : 2.000
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39      1st Qu.: 5.000
##  Median :0.0000   Median :3.000   Median :1.000   Gd:586      Median : 6.000
##  Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735      Mean   : 6.518
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000               3rd Qu.: 7.000
##  Max.   :2.0000   Max.   :8.000   Max.   :3.000               Max.   :14.000
##
##  Functional    Fireplaces    FireplaceQu   GarageType   GarageYrBlt
##  Maj1:  14   Min.   :0.000   Ex  : 24    2Types :  6   Min.   :1900
##  Maj2:   5   1st Qu.:0.000   Fa  : 33    Attchd :870   1st Qu.:1961
```

```
##  Min1:  31   Median :1.000   Gd  :380    Basment: 19   Median :1980
##  Min2:  34   Mean   :0.613   Po  : 20    BuiltIn: 88   Mean   :1979
##  Mod :  15   3rd Qu.:1.000   TA  :313    CarPort:  9   3rd Qu.:2002
##  Sev :   1   Max.   :3.000   NA's:690    Detchd :387   Max.   :2010
##  Typ :1360                               NA's   : 81   NA's   :81
##  GarageFinish   GarageCars      GarageArea     GarageQual  GarageCond
##  Fin :352     Min.   :0.000   Min.   :   0.0   Ex  :   3   Ex  :   2
##  RFn :422     1st Qu.:1.000   1st Qu.: 334.5   Fa  :  48   Fa  :  35
##  Unf :605     Median :2.000   Median : 480.0   Gd  :  14   Gd  :   9
##  NA's: 81     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
##
##  PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch      X3SsnPorch
##  N:  90     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  P:  30     1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Y:1340     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
##
```

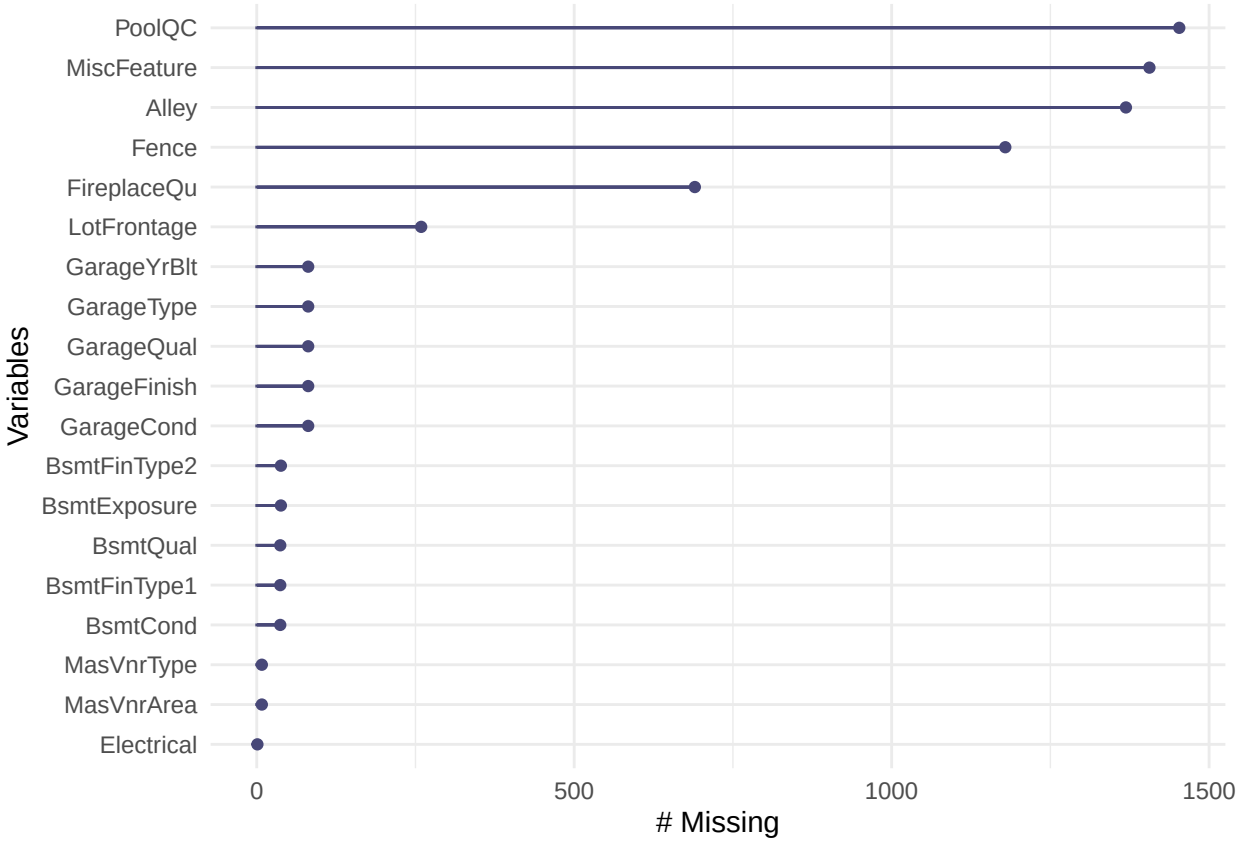
```
## 'data.frame':    1460 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ 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 ...
```

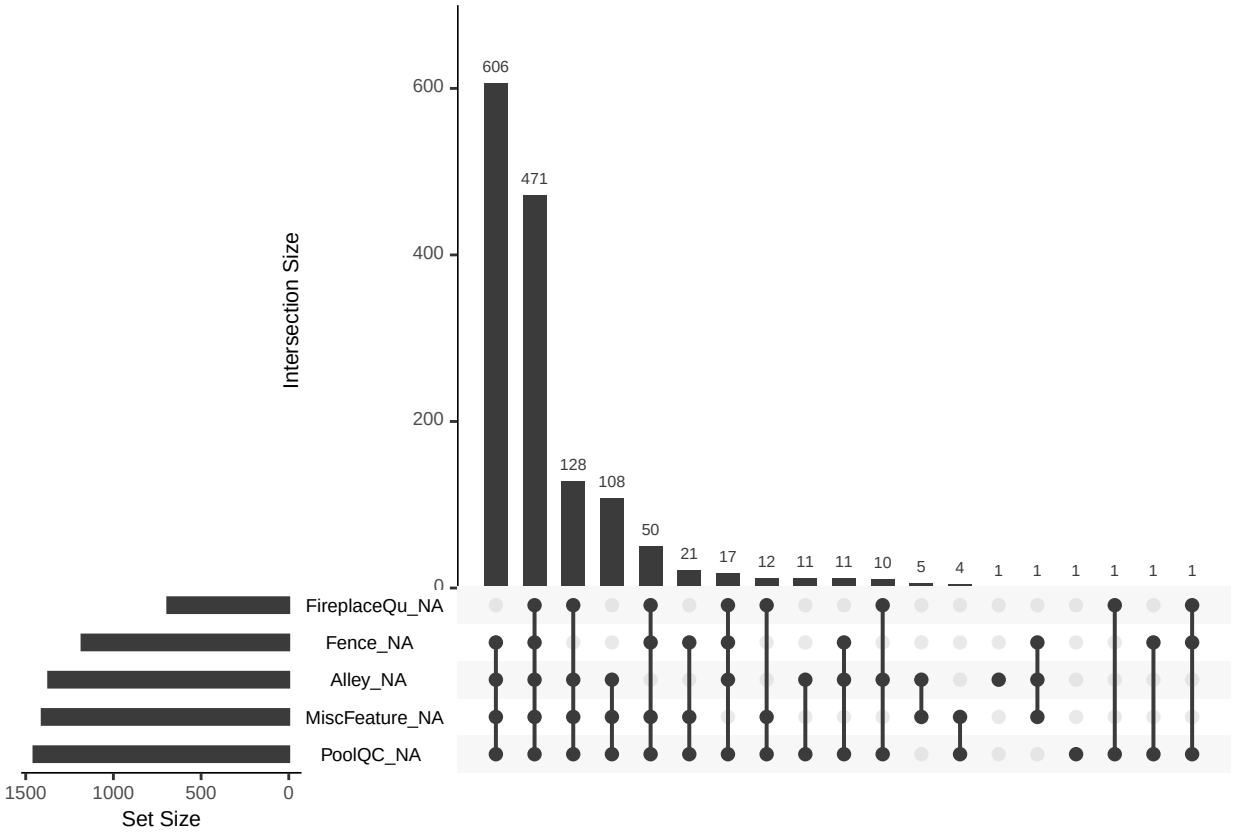
```
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
##  $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 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 ...
```

The dataset consists of 1460 observations and 81 variables, some numeric and some categorical. The target variable has a minimum of 34,950 and a maximum of 7,550,000. The low median compared to the mean suggests some skew.

B. Missing values There are missing values scattered throughout the dataset. We analyse them:





A few categorical features like fireplace, fence, etc. take up the bulk of missings. They do not appear to be important enough to retain so we delete them (FireplaceQu, Fence, Alley, MiscFeature, PoolQC, and LotFrontage). We impute the mean for the rest.

C. Create dummy variables Now we create dummy variables for all of the character variables. Categorical NA's will be handled by adding a dummy variable for NA.

D. Reconcile training and test sets We check if the dataset is missing columns from the test dataset and if so, drop them from the training set. This way we don't risk making predictions on training set variables not found in the test set.

E. Multicollinearity We examine multicollinearity in the dataset. We look at all of the pairs of correlations over .8 There are 24 pairs.

```
##                  col1                  col2 correlation
## 1          TotalBsmtSF             X1stFlrSF   0.8195300
```

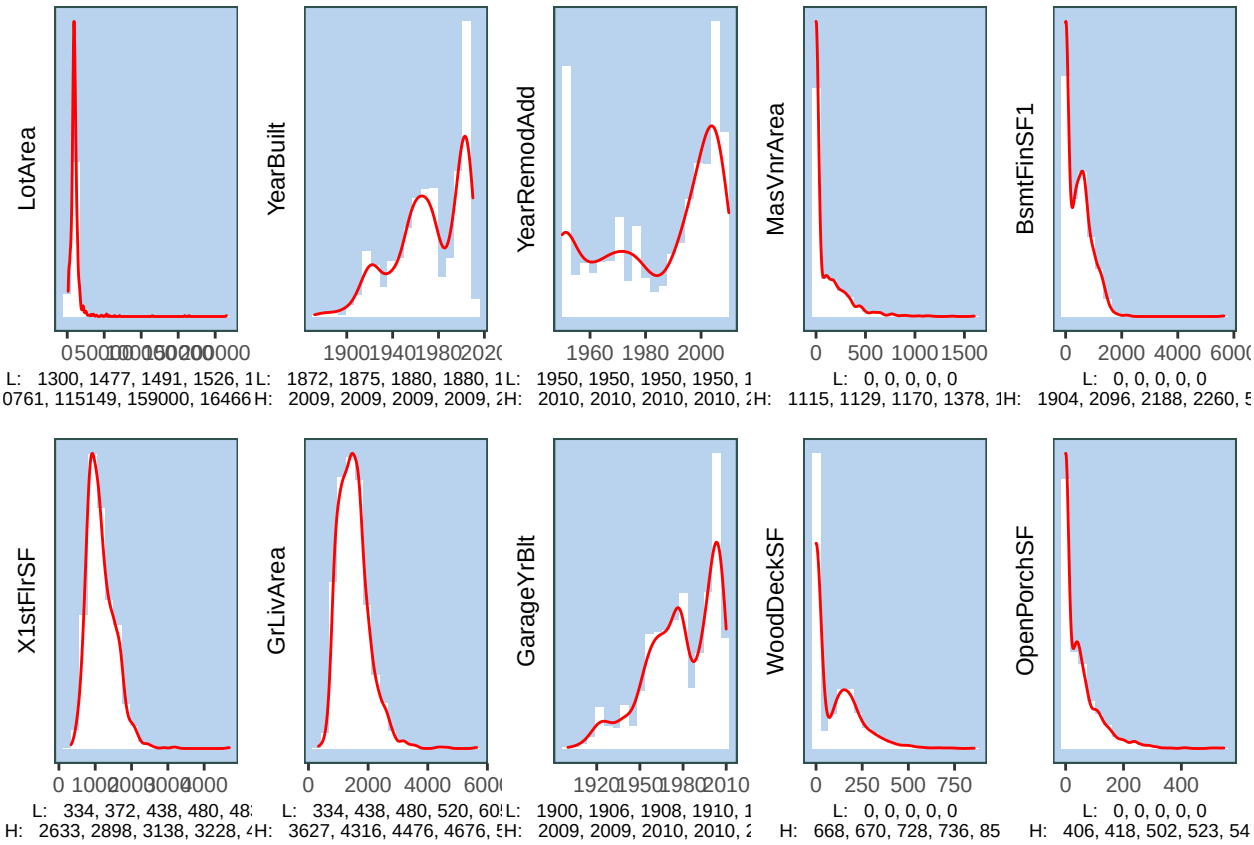
```
## 3            GrLivArea          TotRmsAbvGrd   0.8254894
## 5          GarageCars            GarageArea   0.8824754
## 7         MSZoning_FV  Neighborhood_Somerst   0.8628071
## 9      RoofStyle_Flat      RoofMatl_Tar.Grv   0.8349139
## 11 Exterior1st_AsbShng   Exterior2nd_AsbShng   0.8479167
## 12 Exterior1st_CemntBd   Exterior2nd_CmentBd   0.9741711
## 13 Exterior1st_HdBoard   Exterior2nd_HdBoard   0.8832714
## 14 Exterior1st_MetalSd   Exterior2nd_MetalSd   0.9730652
## 15 Exterior1st_Wd.Sdng   Exterior2nd_Wd.Sdng   0.8592439
## 21     Foundation_Slab           BsmtQual_NA   0.8017334
## 22     Foundation_Slab           BsmtCond_NA   0.8017334
## 23     Foundation_Slab       BsmtFinType1_NA   0.8017334
## 25         BsmtQual_NA           BsmtCond_NA   1.0000000
## 26         BsmtQual_NA       BsmtExposure_NA   0.9864076
## 27         BsmtQual_NA       BsmtFinType1_NA   1.0000000
## 28         BsmtQual_NA       BsmtFinType2_NA   0.9864076
## 31         BsmtCond_NA       BsmtExposure_NA   0.9864076
## 32         BsmtCond_NA       BsmtFinType1_NA   1.0000000
## 33         BsmtCond_NA       BsmtFinType2_NA   0.9864076
## 36     BsmtExposure_NA       BsmtFinType1_NA   0.9864076
## 37     BsmtExposure_NA       BsmtFinType2_NA   0.9729810
## 42     BsmtFinType1_NA       BsmtFinType2_NA   0.9864076
## 47        SaleType_New SaleCondition_Partial   0.9868190
```

Most of the pairs make sense - siding on the first floor will match siding on the second floor, the number of cars a garage can hold will be related to its area. We will address the multicollinearity more closely when we run the analysis.

2. Transformations

A. Log of SalePrice The skew in the dependent variable suggests a log transformation.

B. Other transformations A number of histograms suggest issues with some of the independent variables.



We can see some transformations might be useful. We: 1. Add a dummy variable to mark YearBuilt before and after 1920 2. We set YearRemodAdd = 1950 to 0, and create a dummy variable YearRemodUnknown to track it 3. We add dummies for NoFinBsmt, HasDeck, and HasPorch 4. We eliminate outliers by setting GrLivArea<4000

3. Model and Predict:

A. Base Model We run an OLS regression using the stepAIC algorithm to minimize AIC.

```
##
## Call:
## lm(formula = SalePrice ~ Id + LotArea + OverallQual + OverallCond +
##     YearBuilt + YearRemodAdd + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +
##     X1stFlrSF + X2ndFlrSF + LowQualFinSF + BsmtFullBath + FullBath +
##     HalfBath + KitchenAbvGr + Fireplaces + GarageCars + GarageArea +
##     WoodDeckSF + OpenPorchSF + EnclosedPorch + X3SsnPorch + ScreenPorch +
```

```
##     PoolArea + MSZoning_C..all. + MSZoning_RM + Street_Grvl +
##     LotConfig_CulDSac + LandSlope_Sev + Neighborhood_BrkSide +
##     Neighborhood_ClearCr + Neighborhood_Crawfor + Neighborhood_Edwards +
##     Neighborhood_IDOTRR + Neighborhood_MeadowV + Neighborhood_Mitchel +
##     Neighborhood_NWAmes + Neighborhood_NoRidge + Neighborhood_NridgHt +
##     Neighborhood_Somerst + Neighborhood_StoneBr + Condition1_Artery +
##     Condition1_RRAe + Condition1_RRAn + BldgType_Duplex + BldgType_Twnhs +
##     BldgType_TwnhsE + RoofStyle_Flat + Exterior1st_BrkComm +
##     Exterior1st_BrkFace + Exterior1st_CemntBd + Exterior1st_HdBoard +
##     Exterior1st_Plywood + Exterior1st_Wd.Sdng + Exterior2nd_CmentBd +
##     Exterior2nd_Wd.Sdng + MasVnrType_BrkCmn + MasVnrType_Stone +
##     ExterCond_Ex + ExterCond_Fa + ExterCond_Gd + Foundation_Stone +
##     Foundation_Wood + BsmtQual_Ex + BsmtCond_Fa + BsmtCond_Po +
##     BsmtExposure_Gd + BsmtFinType2_BLQ + BsmtFinType2_GLQ + Heating_GasW +
##     Heating_Grav + Heating_Wall + CentralAir_N + KitchenQual_Ex +
##     Functional_Maj1 + Functional_Maj2 + Functional_Min1 + Functional_Min2 +
##     Functional_Mod + Functional_Sev + GarageType_2Types + GarageType_NA +
##     GarageQual_Fa + GarageQual_Po + GarageCond_Fa + GarageCond_Po +
##     PavedDrive_N + SaleType_CWD + SaleType_Con + SaleType_ConLD +
##     SaleType_ConLw + SaleType_New + SaleType_Oth + SaleCondition_Abnorml +
##     SaleCondition_Family + SaleCondition_Partial + BuiltAfter1920 +
##     YearRemodUnknown, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67671 -0.04919  0.00285  0.05111  0.46324
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.497e+00  5.680e-01   7.918 4.97e-15 ***
## Id                    -9.095e-06  6.429e-06  -1.415 0.157355
## LotArea                2.600e-06  3.859e-07   6.738 2.37e-11 ***
## OverallQual            4.738e-02  3.865e-03  12.258  < 2e-16 ***
```

```
## OverallCond            3.739e-02  3.501e-03  10.681  < 2e-16 ***
## YearBuilt              2.426e-03  2.618e-04   9.267  < 2e-16 ***
## YearRemodAdd           8.239e-04  2.532e-04   3.255 0.001164 **
## BsmtFinSF1             1.614e-04  1.476e-05  10.936  < 2e-16 ***
## BsmtFinSF2             1.153e-04  2.312e-05   4.986 6.98e-07 ***
## BsmtUnfSF              9.846e-05  1.305e-05   7.548 8.09e-14 ***
## X1stFlrSF              2.669e-04  1.638e-05  16.293  < 2e-16 ***
## X2ndFlrSF              2.570e-04  1.274e-05  20.169  < 2e-16 ***
## LowQualFinSF           2.394e-04  5.854e-05   4.090 4.58e-05 ***
## BsmtFullBath           2.752e-02  7.453e-03   3.693 0.000231 ***
## FullBath               1.898e-02  8.570e-03   2.215 0.026912 *
## HalfBath               1.867e-02  8.196e-03   2.277 0.022926 *
## KitchenAbvGr          -5.083e-02  1.895e-02  -2.683 0.007388 **
## Fireplaces             1.974e-02  5.418e-03   3.643 0.000279 ***
## GarageCars             3.053e-02  9.111e-03   3.351 0.000827 ***
## GarageArea             6.558e-05  3.021e-05   2.171 0.030083 *
## WoodDeckSF             8.737e-05  2.401e-05   3.639 0.000284 ***
## OpenPorchSF            1.055e-04  4.660e-05   2.264 0.023702 *
## EnclosedPorch          1.094e-04  5.013e-05   2.182 0.029263 *
## X3SsnPorch             1.885e-04  9.255e-05   2.037 0.041878 *
## ScreenPorch            2.295e-04  5.078e-05   4.518 6.78e-06 ***
## PoolArea               1.175e-04  7.820e-05   1.503 0.133101
## MSZoning_C..all.      -4.260e-01  4.027e-02 -10.580  < 2e-16 ***
## MSZoning_RM           -6.101e-02  1.060e-02  -5.756 1.06e-08 ***
## Street_Grvl           -8.150e-02  4.677e-02  -1.742 0.081684 .
## LotConfig_CulDSac      3.232e-02  1.157e-02   2.793 0.005289 **
## LandSlope_Sev         -1.274e-01  3.958e-02  -3.218 0.001323 **
## Neighborhood_BrkSide   5.767e-02  1.588e-02   3.631 0.000293 ***
## Neighborhood_ClearCr   5.267e-02  2.233e-02   2.358 0.018517 *
## Neighborhood_Crawfor   1.453e-01  1.659e-02   8.760  < 2e-16 ***
## Neighborhood_Edwards  -3.461e-02  1.193e-02  -2.901 0.003782 **
## Neighborhood_IDOTRR    3.522e-02  2.210e-02   1.593 0.111325
## Neighborhood_MeadowV  -1.180e-01  3.263e-02  -3.617 0.000309 ***
```

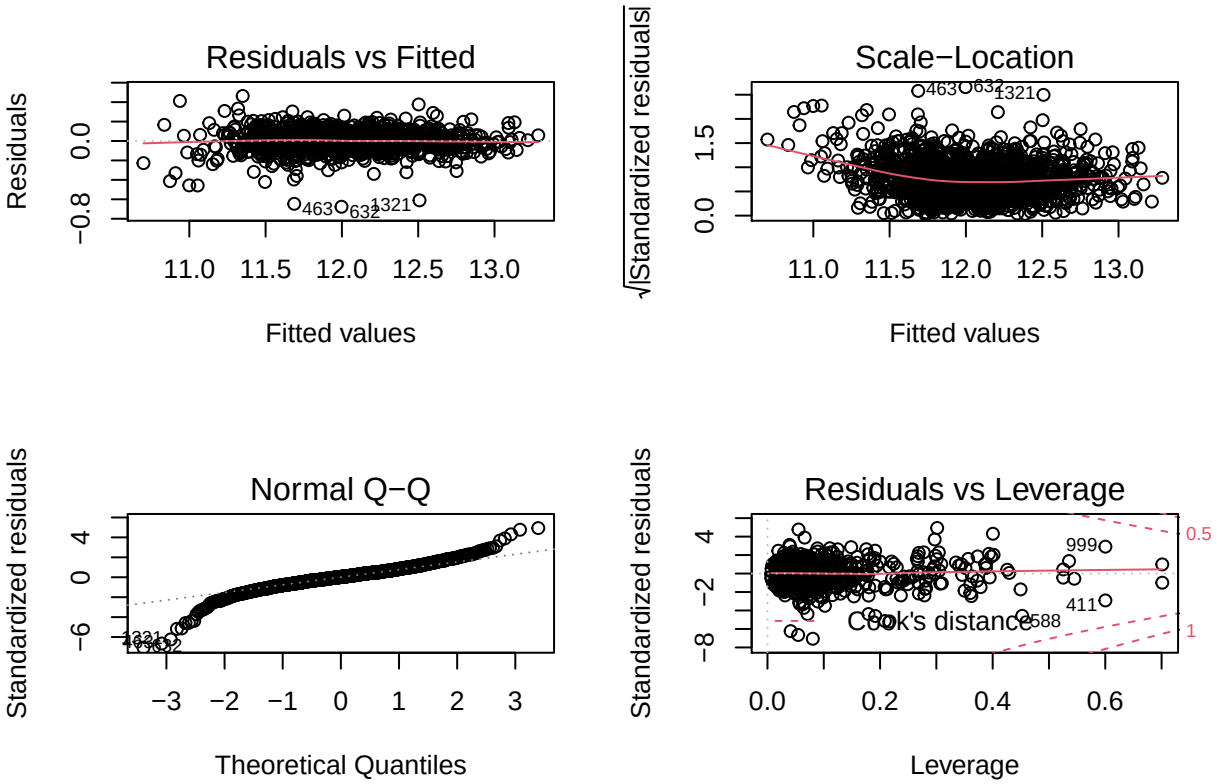
```
## Neighborhood_Mitchel  -2.182e-02  1.573e-02  -1.387 0.165621
## Neighborhood_NWAmes   -2.698e-02  1.351e-02  -1.997 0.046026 *
## Neighborhood_NoRidge   5.057e-02  1.853e-02   2.729 0.006443 **
## Neighborhood_NridgHt   6.080e-02  1.605e-02   3.789 0.000158 ***
## Neighborhood_Somerst   4.679e-02  1.377e-02   3.399 0.000696 ***
## Neighborhood_StoneBr   1.244e-01  2.282e-02   5.449 6.00e-08 ***
## Condition1_Artery     -6.968e-02  1.609e-02  -4.331 1.60e-05 ***
## Condition1_RRAe       -1.178e-01  3.122e-02  -3.774 0.000168 ***
## Condition1_RRAn       -3.755e-02  2.077e-02  -1.808 0.070808 .
## BldgType_Duplex       -4.643e-02  2.187e-02  -2.123 0.033904 *
## BldgType_Twnhs        -9.935e-02  1.853e-02  -5.362 9.68e-08 ***
## BldgType_TwnhsE       -3.713e-02  1.224e-02  -3.034 0.002460 **
## RoofStyle_Flat         6.984e-02  3.551e-02   1.967 0.049422 *
## Exterior1st_BrkComm   -2.137e-01  7.879e-02  -2.712 0.006775 **
## Exterior1st_BrkFace    3.442e-02  1.680e-02   2.049 0.040636 *
## Exterior1st_CemntBd   -9.097e-02  6.111e-02  -1.489 0.136825
## Exterior1st_HdBoard   -3.707e-02  8.732e-03  -4.246 2.33e-05 ***
## Exterior1st_Plywood   -3.344e-02  1.206e-02  -2.772 0.005649 **
## Exterior1st_Wd.Sdng   -6.159e-02  1.661e-02  -3.707 0.000218 ***
## Exterior2nd_CmentBd    8.753e-02  6.156e-02   1.422 0.155343
## Exterior2nd_Wd.Sdng    4.423e-02  1.637e-02   2.702 0.006970 **
## MasVnrType_BrkCmn     -5.277e-02  2.780e-02  -1.898 0.057862 .
## MasVnrType_Stone       2.179e-02  1.121e-02   1.944 0.052089 .
## ExterCond_Ex           1.013e-01  6.055e-02   1.672 0.094696 .
## ExterCond_Fa          -3.659e-02  2.264e-02  -1.616 0.106303
## ExterCond_Gd          -1.809e-02  9.700e-03  -1.864 0.062477 .
## Foundation_Stone       1.026e-01  4.444e-02   2.309 0.021089 *
## Foundation_Wood       -1.391e-01  5.955e-02  -2.336 0.019641 *
## BsmtQual_Ex            2.848e-02  1.331e-02   2.140 0.032529 *
## BsmtCond_Fa           -3.251e-02  1.692e-02  -1.922 0.054858 .
## BsmtCond_Po            1.691e-01  9.088e-02   1.861 0.062987 .
## BsmtExposure_Gd        5.052e-02  1.093e-02   4.621 4.19e-06 ***
## BsmtFinType2_BLQ      -3.235e-02  1.922e-02  -1.683 0.092602 .
```

```
## BsmtFinType2_GLQ       5.425e-02  3.079e-02   1.762 0.078250 .
## Heating_GasW           5.224e-02  2.727e-02   1.916 0.055622 .
## Heating_Grav          -1.477e-01  4.379e-02  -3.372 0.000766 ***
## Heating_Wall           8.911e-02  5.559e-02   1.603 0.109163
## CentralAir_N          -6.317e-02  1.427e-02  -4.426 1.04e-05 ***
## KitchenQual_Ex         6.301e-02  1.374e-02   4.587 4.92e-06 ***
## Functional_Maj1       -9.671e-02  2.996e-02  -3.228 0.001278 **
## Functional_Maj2       -3.104e-01  4.978e-02  -6.237 5.96e-10 ***
## Functional_Min1       -4.847e-02  1.949e-02  -2.487 0.013004 *
## Functional_Min2       -3.440e-02  1.879e-02  -1.830 0.067441 .
## Functional_Mod        -1.175e-01  2.820e-02  -4.167 3.28e-05 ***
## Functional_Sev        -4.125e-01  1.119e-01  -3.685 0.000238 ***
## GarageType_2Types     -8.141e-02  4.455e-02  -1.827 0.067892 .
## GarageType_NA         -4.205e-02  1.710e-02  -2.459 0.014066 *
## GarageQual_Fa         -3.572e-02  1.914e-02  -1.866 0.062228 .
## GarageQual_Po         -1.359e-01  8.849e-02  -1.536 0.124782
## GarageCond_Fa         -3.201e-02  2.103e-02  -1.522 0.128259
## GarageCond_Po          1.079e-01  5.623e-02   1.919 0.055189 .
## PavedDrive_N          -1.862e-02  1.352e-02  -1.378 0.168529
## SaleType_CWD           8.151e-02  5.230e-02   1.559 0.119343
## SaleType_Con           1.233e-01  7.211e-02   1.710 0.087502 .
## SaleType_ConLD         1.333e-01  3.657e-02   3.645 0.000277 ***
## SaleType_ConLw         6.869e-02  4.704e-02   1.460 0.144458
## SaleType_New           1.772e-01  6.258e-02   2.832 0.004696 **
## SaleType_Oth           8.899e-02  5.994e-02   1.485 0.137864
## SaleCondition_Abnorml -6.940e-02  1.128e-02  -6.150 1.02e-09 ***
## SaleCondition_Family  -4.724e-02  2.339e-02  -2.020 0.043582 *
## SaleCondition_Partial -1.311e-01  6.197e-02  -2.116 0.034563 *
## BuiltAfter1920        -2.748e-02  1.554e-02  -1.768 0.077271 .
## YearRemodUnknown       1.598e+00  5.011e-01   3.188 0.001464 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.09995 on 1356 degrees of freedom
## Multiple R-squared:  0.9406, Adjusted R-squared:  0.9363
## F-statistic: 217.1 on 99 and 1356 DF,  p-value: < 2.2e-16
##
## [1] "VIF Analysis"
##                     Id               LotArea           OverallQual
##               1.070446              2.109227              4.082242
##            OverallCond             YearBuilt          YearRemodAdd
##               2.214804              9.101482           3969.919813
##             BsmtFinSF1            BsmtFinSF2             BsmtUnfSF
##               5.870162              2.030635              4.846297
##              X1stFlrSF             X2ndFlrSF          LowQualFinSF
##               5.329108              4.402165              1.183029
##          BsmtFullBath              FullBath              HalfBath
##               2.165471              3.208325              2.470764
##          KitchenAbvGr            Fireplaces            GarageCars
##               2.544832              1.752755              6.732297
##            GarageArea            WoodDeckSF           OpenPorchSF
##               5.970826              1.315971              1.350534
##         EnclosedPorch            X3SsnPorch           ScreenPorch
##               1.370222              1.074979              1.170604
##              PoolArea      MSZoning_C..all.           MSZoning_RM
##               1.114980              1.611695              2.083850
##           Street_Grvl     LotConfig_CulDSac         LandSlope_Sev
##               1.308506              1.178051              2.020225
##  Neighborhood_BrkSide  Neighborhood_ClearCr  Neighborhood_Crawfor
##               1.406034              1.371166              1.355105
##  Neighborhood_Edwards   Neighborhood_IDOTRR  Neighborhood_MeadowV
##               1.302697              1.763348              1.790537
##  Neighborhood_Mitchel   Neighborhood_NWAmes  Neighborhood_NoRidge
##               1.172639              1.266834              1.305002
##  Neighborhood_NridgHt  Neighborhood_Somerst  Neighborhood_StoneBr
##               1.879891              1.535288              1.281182
```

```
##     Condition1_Artery       Condition1_RRAe       Condition1_RRAn
##              1.202671              1.065123              1.102368
##       BldgType_Duplex       BldgType_Twnhs      BldgType_TwnhsE
##              2.399890              1.434197              1.575527
##       RoofStyle_Flat   Exterior1st_BrkComm   Exterior1st_BrkFace
##              1.626001              1.240871              1.363344
##   Exterior1st_CemntBd   Exterior1st_HdBoard   Exterior1st_Plywood
##             21.501494              1.430582              1.456569
##   Exterior1st_Wd.Sdng   Exterior2nd_CmentBd   Exterior2nd_Wd.Sdng
##              4.865587             21.475610              4.567847
##     MasVnrType_BrkCmn     MasVnrType_Stone         ExterCond_Ex
##              1.148006              1.447534              1.098791
##         ExterCond_Fa         ExterCond_Gd     Foundation_Stone
##              1.408762              1.237170              1.181241
##       Foundation_Wood          BsmtQual_Ex          BsmtCond_Fa
##              1.062645              1.907796              1.249316
##          BsmtCond_Po      BsmtExposure_Gd     BsmtFinType2_BLQ
##              1.651124              1.426619              1.192674
##     BsmtFinType2_GLQ         Heating_GasW         Heating_Grav
##              1.315372              1.323281              1.336804
##         Heating_Wall         CentralAir_N       KitchenQual_Ex
##              1.233841              1.811117              1.693583
##       Functional_Maj1       Functional_Maj2       Functional_Min1
##              1.245837              1.235746              1.153368
##       Functional_Min2       Functional_Mod       Functional_Sev
##              1.173951              1.181688              1.253489
##    GarageType_2Types        GarageType_NA        GarageQual_Fa
##              1.187210              2.238917              1.702411
##        GarageQual_Po        GarageCond_Fa        GarageCond_Po
##              2.346550              1.512048              2.204323
##         PavedDrive_N         SaleType_CWD         SaleType_Con
##              1.543691              1.092156              1.039413
##        SaleType_ConLD       SaleType_ConLw         SaleType_New
```

```
##              1.197037              1.103716             43.157779
##          SaleType_Oth SaleCondition_Abnorml  SaleCondition_Family
##              1.076456              1.186853              1.079773
## SaleCondition_Partial       BuiltAfter1920      YearRemodUnknown
##             43.285823              2.622103           3926.926004
```




```
## NULL
##
##  studentized Breusch-Pagan test
##
## data:  step3
## BP = 270.14, df = 99, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
```

```
## data:  step3$residuals
## W = 0.93874, p-value < 2.2e-16
##
## [1] "AIC:  -2476.10190734742"

##
## Call:
## lm(formula = SalePrice ~ Id + LotArea + OverallQual + OverallCond +
##     YearBuilt + YearRemodAdd + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +
##     X1stFlrSF + X2ndFlrSF + LowQualFinSF + BsmtFullBath + FullBath +
##     HalfBath + KitchenAbvGr + Fireplaces + GarageCars + GarageArea +
##     WoodDeckSF + OpenPorchSF + EnclosedPorch + X3SsnPorch + ScreenPorch +
##     PoolArea + MSZoning_C..all. + MSZoning_RM + Street_Grvl +
##     LotConfig_CulDSac + LandSlope_Sev + Neighborhood_BrkSide +
##     Neighborhood_ClearCr + Neighborhood_Crawfor + Neighborhood_Edwards +
##     Neighborhood_IDOTRR + Neighborhood_MeadowV + Neighborhood_Mitchel +
##     Neighborhood_NWAmes + Neighborhood_NoRidge + Neighborhood_NridgHt +
##     Neighborhood_Somerst + Neighborhood_StoneBr + Condition1_Artery +
##     Condition1_RRAe + Condition1_RRAn + BldgType_Duplex + BldgType_Twnhs +
##     BldgType_TwnhsE + RoofStyle_Flat + Exterior1st_BrkComm +
##     Exterior1st_BrkFace + Exterior1st_CemntBd + Exterior1st_HdBoard +
##     Exterior1st_Plywood + Exterior1st_Wd.Sdng + Exterior2nd_CmentBd +
##     Exterior2nd_Wd.Sdng + MasVnrType_BrkCmn + MasVnrType_Stone +
##     ExterCond_Ex + ExterCond_Fa + ExterCond_Gd + Foundation_Stone +
##     Foundation_Wood + BsmtQual_Ex + BsmtCond_Fa + BsmtCond_Po +
##     BsmtExposure_Gd + BsmtFinType2_BLQ + BsmtFinType2_GLQ + Heating_GasW +
##     Heating_Grav + Heating_Wall + CentralAir_N + KitchenQual_Ex +
##     Functional_Maj1 + Functional_Maj2 + Functional_Min1 + Functional_Min2 +
##     Functional_Mod + Functional_Sev + GarageType_2Types + GarageType_NA +
##     GarageQual_Fa + GarageQual_Po + GarageCond_Fa + GarageCond_Po +
##     PavedDrive_N + SaleType_CWD + SaleType_Con + SaleType_ConLD +
##     SaleType_ConLw + SaleType_New + SaleType_Oth + SaleCondition_Abnorml +
##     SaleCondition_Family + SaleCondition_Partial + BuiltAfter1920 +
```

```
##     YearRemodUnknown, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67671 -0.04919  0.00285  0.05111  0.46324
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.497e+00  5.680e-01   7.918 4.97e-15 ***
## Id                    -9.095e-06  6.429e-06  -1.415 0.157355
## LotArea                2.600e-06  3.859e-07   6.738 2.37e-11 ***
## OverallQual            4.738e-02  3.865e-03  12.258  < 2e-16 ***
## OverallCond            3.739e-02  3.501e-03  10.681  < 2e-16 ***
## YearBuilt              2.426e-03  2.618e-04   9.267  < 2e-16 ***
## YearRemodAdd           8.239e-04  2.532e-04   3.255 0.001164 **
## BsmtFinSF1             1.614e-04  1.476e-05  10.936  < 2e-16 ***
## BsmtFinSF2             1.153e-04  2.312e-05   4.986 6.98e-07 ***
## BsmtUnfSF              9.846e-05  1.305e-05   7.548 8.09e-14 ***
## X1stFlrSF              2.669e-04  1.638e-05  16.293  < 2e-16 ***
## X2ndFlrSF              2.570e-04  1.274e-05  20.169  < 2e-16 ***
## LowQualFinSF           2.394e-04  5.854e-05   4.090 4.58e-05 ***
## BsmtFullBath           2.752e-02  7.453e-03   3.693 0.000231 ***
## FullBath               1.898e-02  8.570e-03   2.215 0.026912 *
## HalfBath               1.867e-02  8.196e-03   2.277 0.022926 *
## KitchenAbvGr          -5.083e-02  1.895e-02  -2.683 0.007388 **
## Fireplaces             1.974e-02  5.418e-03   3.643 0.000279 ***
## GarageCars             3.053e-02  9.111e-03   3.351 0.000827 ***
## GarageArea             6.558e-05  3.021e-05   2.171 0.030083 *
## WoodDeckSF             8.737e-05  2.401e-05   3.639 0.000284 ***
## OpenPorchSF            1.055e-04  4.660e-05   2.264 0.023702 *
## EnclosedPorch          1.094e-04  5.013e-05   2.182 0.029263 *
## X3SsnPorch             1.885e-04  9.255e-05   2.037 0.041878 *
## ScreenPorch            2.295e-04  5.078e-05   4.518 6.78e-06 ***
```

```
## PoolArea               1.175e-04  7.820e-05   1.503 0.133101
## MSZoning_C..all.      -4.260e-01  4.027e-02 -10.580  < 2e-16 ***
## MSZoning_RM           -6.101e-02  1.060e-02  -5.756 1.06e-08 ***
## Street_Grvl           -8.150e-02  4.677e-02  -1.742 0.081684 .
## LotConfig_CulDSac      3.232e-02  1.157e-02   2.793 0.005289 **
## LandSlope_Sev         -1.274e-01  3.958e-02  -3.218 0.001323 **
## Neighborhood_BrkSide   5.767e-02  1.588e-02   3.631 0.000293 ***
## Neighborhood_ClearCr   5.267e-02  2.233e-02   2.358 0.018517 *
## Neighborhood_Crawfor   1.453e-01  1.659e-02   8.760  < 2e-16 ***
## Neighborhood_Edwards  -3.461e-02  1.193e-02  -2.901 0.003782 **
## Neighborhood_IDOTRR    3.522e-02  2.210e-02   1.593 0.111325
## Neighborhood_MeadowV  -1.180e-01  3.263e-02  -3.617 0.000309 ***
## Neighborhood_Mitchel  -2.182e-02  1.573e-02  -1.387 0.165621
## Neighborhood_NWAmes   -2.698e-02  1.351e-02  -1.997 0.046026 *
## Neighborhood_NoRidge   5.057e-02  1.853e-02   2.729 0.006443 **
## Neighborhood_NridgHt   6.080e-02  1.605e-02   3.789 0.000158 ***
## Neighborhood_Somerst   4.679e-02  1.377e-02   3.399 0.000696 ***
## Neighborhood_StoneBr   1.244e-01  2.282e-02   5.449 6.00e-08 ***
## Condition1_Artery     -6.968e-02  1.609e-02  -4.331 1.60e-05 ***
## Condition1_RRAe       -1.178e-01  3.122e-02  -3.774 0.000168 ***
## Condition1_RRAn       -3.755e-02  2.077e-02  -1.808 0.070808 .
## BldgType_Duplex       -4.643e-02  2.187e-02  -2.123 0.033904 *
## BldgType_Twnhs        -9.935e-02  1.853e-02  -5.362 9.68e-08 ***
## BldgType_TwnhsE       -3.713e-02  1.224e-02  -3.034 0.002460 **
## RoofStyle_Flat         6.984e-02  3.551e-02   1.967 0.049422 *
## Exterior1st_BrkComm   -2.137e-01  7.879e-02  -2.712 0.006775 **
## Exterior1st_BrkFace    3.442e-02  1.680e-02   2.049 0.040636 *
## Exterior1st_CemntBd   -9.097e-02  6.111e-02  -1.489 0.136825
## Exterior1st_HdBoard   -3.707e-02  8.732e-03  -4.246 2.33e-05 ***
## Exterior1st_Plywood   -3.344e-02  1.206e-02  -2.772 0.005649 **
## Exterior1st_Wd.Sdng   -6.159e-02  1.661e-02  -3.707 0.000218 ***
## Exterior2nd_CmentBd    8.753e-02  6.156e-02   1.422 0.155343
## Exterior2nd_Wd.Sdng    4.423e-02  1.637e-02   2.702 0.006970 **
```

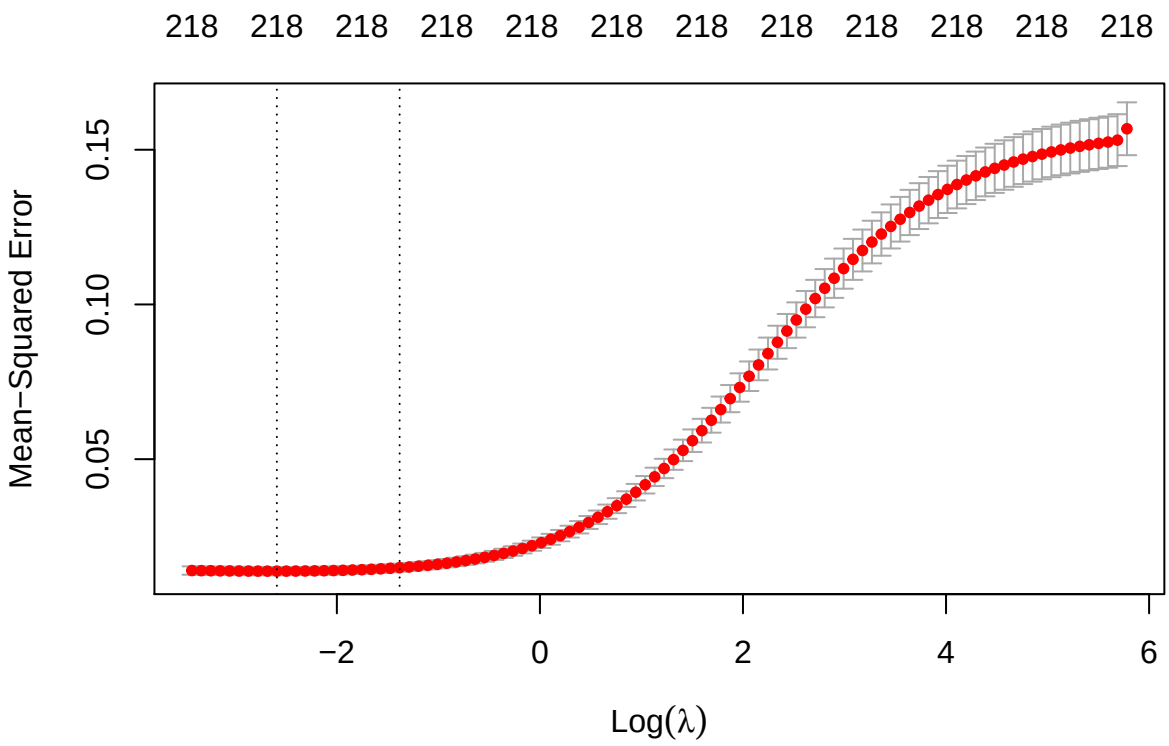
```
## MasVnrType_BrkCmn     -5.277e-02  2.780e-02  -1.898 0.057862 .
## MasVnrType_Stone       2.179e-02  1.121e-02   1.944 0.052089 .
## ExterCond_Ex           1.013e-01  6.055e-02   1.672 0.094696 .
## ExterCond_Fa          -3.659e-02  2.264e-02  -1.616 0.106303
## ExterCond_Gd          -1.809e-02  9.700e-03  -1.864 0.062477 .
## Foundation_Stone       1.026e-01  4.444e-02   2.309 0.021089 *
## Foundation_Wood       -1.391e-01  5.955e-02  -2.336 0.019641 *
## BsmtQual_Ex            2.848e-02  1.331e-02   2.140 0.032529 *
## BsmtCond_Fa           -3.251e-02  1.692e-02  -1.922 0.054858 .
## BsmtCond_Po            1.691e-01  9.088e-02   1.861 0.062987 .
## BsmtExposure_Gd        5.052e-02  1.093e-02   4.621 4.19e-06 ***
## BsmtFinType2_BLQ      -3.235e-02  1.922e-02  -1.683 0.092602 .
## BsmtFinType2_GLQ       5.425e-02  3.079e-02   1.762 0.078250 .
## Heating_GasW           5.224e-02  2.727e-02   1.916 0.055622 .
## Heating_Grav          -1.477e-01  4.379e-02  -3.372 0.000766 ***
## Heating_Wall           8.911e-02  5.559e-02   1.603 0.109163
## CentralAir_N          -6.317e-02  1.427e-02  -4.426 1.04e-05 ***
## KitchenQual_Ex         6.301e-02  1.374e-02   4.587 4.92e-06 ***
## Functional_Maj1       -9.671e-02  2.996e-02  -3.228 0.001278 **
## Functional_Maj2       -3.104e-01  4.978e-02  -6.237 5.96e-10 ***
## Functional_Min1       -4.847e-02  1.949e-02  -2.487 0.013004 *
## Functional_Min2       -3.440e-02  1.879e-02  -1.830 0.067441 .
## Functional_Mod        -1.175e-01  2.820e-02  -4.167 3.28e-05 ***
## Functional_Sev        -4.125e-01  1.119e-01  -3.685 0.000238 ***
## GarageType_2Types     -8.141e-02  4.455e-02  -1.827 0.067892 .
## GarageType_NA         -4.205e-02  1.710e-02  -2.459 0.014066 *
## GarageQual_Fa         -3.572e-02  1.914e-02  -1.866 0.062228 .
## GarageQual_Po         -1.359e-01  8.849e-02  -1.536 0.124782
## GarageCond_Fa         -3.201e-02  2.103e-02  -1.522 0.128259
## GarageCond_Po          1.079e-01  5.623e-02   1.919 0.055189 .
## PavedDrive_N          -1.862e-02  1.352e-02  -1.378 0.168529
## SaleType_CWD           8.151e-02  5.230e-02   1.559 0.119343
## SaleType_Con           1.233e-01  7.211e-02   1.710 0.087502 .
```

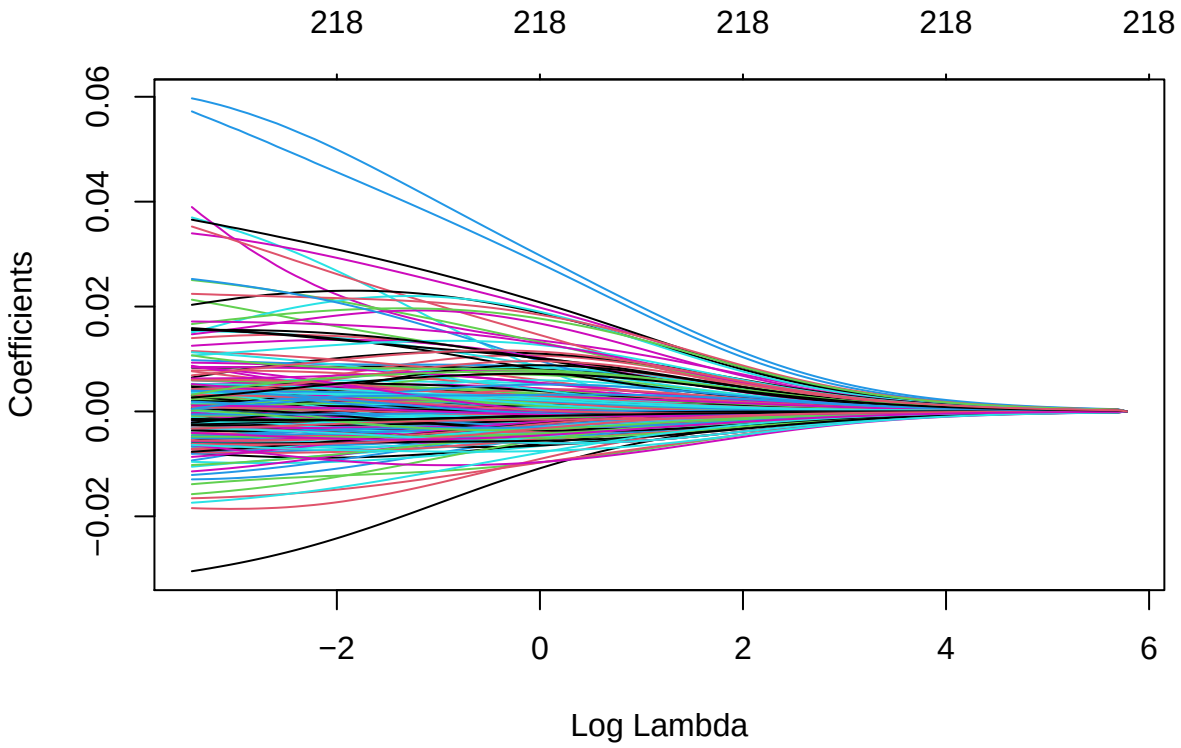
```
## SaleType_ConLD         1.333e-01  3.657e-02   3.645 0.000277 ***
## SaleType_ConLw         6.869e-02  4.704e-02   1.460 0.144458
## SaleType_New           1.772e-01  6.258e-02   2.832 0.004696 **
## SaleType_Oth           8.899e-02  5.994e-02   1.485 0.137864
## SaleCondition_Abnorml -6.940e-02  1.128e-02  -6.150 1.02e-09 ***
## SaleCondition_Family  -4.724e-02  2.339e-02  -2.020 0.043582 *
## SaleCondition_Partial -1.311e-01  6.197e-02  -2.116 0.034563 *
## BuiltAfter1920        -2.748e-02  1.554e-02  -1.768 0.077271 .
## YearRemodUnknown       1.598e+00  5.011e-01   3.188 0.001464 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09995 on 1356 degrees of freedom
## Multiple R-squared:  0.9406, Adjusted R-squared:  0.9363
## F-statistic: 217.1 on 99 and 1356 DF,  p-value: < 2.2e-16
```

Now we make predictions

We achieve a score of .14586 on kaggle. This puts us in the 60th percentile.

B. Ridge regression: R makes it easy to find the best lambda by using kfold validation. Below are the results of our ridge regression analysis. Unlike stepAIC, ridge regression will retain all of the variables.


218 218 218 218 218 218 218 218 218 218 218 218
Mean−Squared Error
0.05
0.10
0.15
−2
0
2
4
6
Log(λ)




```
## 219 x 1 sparse Matrix of class "dgCMatrix"
##                                    s0
## (Intercept)             1.202194e+01
## Id                     -3.261232e-03
## MSSubClass             -1.279049e-03
## LotArea                 1.833604e-02
## OverallQual             5.483452e-02
## OverallCond             3.170589e-02
## YearBuilt               2.764065e-02
## YearRemodAdd            8.324423e-03
## MasVnrArea              5.065497e-03
## BsmtFinSF1              2.283311e-02
## BsmtFinSF2              3.214778e-03
## BsmtUnfSF               5.951193e-03
## TotalBsmtSF             3.140363e-02
```

```
## X1stFlrSF              3.328663e-02
## X2ndFlrSF              2.967100e-02
## LowQualFinSF           1.457690e-03
## GrLivArea              5.062906e-02
## BsmtFullBath           1.484172e-02
## BsmtHalfBath           6.738830e-04
## FullBath               2.244747e-02
## HalfBath               1.462801e-02
## BedroomAbvGr           2.264154e-03
## KitchenAbvGr          -1.211183e-02
## TotRmsAbvGrd           1.915354e-02
## Fireplaces             1.683929e-02
## GarageYrBlt            8.865600e-03
## GarageCars             2.190916e-02
## GarageArea             1.844499e-02
## WoodDeckSF             9.285274e-03
## OpenPorchSF            5.538145e-03
## EnclosedPorch          5.218463e-03
## X3SsnPorch             5.830328e-03
## ScreenPorch            1.065966e-02
## PoolArea               5.182842e-03
## MiscVal               -1.759561e-03
## MoSold                -6.433307e-04
## YrSold                -1.820294e-03
## MSZoning_C..all.      -2.733442e-02
## MSZoning_FV            7.688990e-03
## MSZoning_RM           -1.277543e-02
## Street_Grvl           -5.453089e-03
## LotShape_IR1           1.471177e-03
## LotShape_IR2           4.726946e-03
## LotShape_IR3           1.442141e-03
## LandContour_Bnk       -1.268544e-03
## LandContour_HLS        3.082793e-03
```

```
## LandContour_Low       -8.679675e-04
## LotConfig_Corner       2.762383e-03
## LotConfig_CulDSac      8.926308e-03
## LotConfig_FR2         -5.065509e-03
## LotConfig_FR3         -1.434172e-03
## LandSlope_Mod          2.207995e-03
## LandSlope_Sev         -6.669501e-03
## Neighborhood_Blmngtn  -1.450177e-04
## Neighborhood_Blueste  -2.881479e-03
## Neighborhood_BrDale   -8.695861e-03
## Neighborhood_BrkSide   5.834388e-03
## Neighborhood_ClearCr   5.602871e-03
## Neighborhood_Crawfor   2.299333e-02
## Neighborhood_Edwards  -9.824952e-03
## Neighborhood_Gilbert   4.973214e-04
## Neighborhood_IDOTRR   -3.059494e-03
## Neighborhood_MeadowV  -1.838445e-02
## Neighborhood_Mitchel  -4.536545e-03
## Neighborhood_NPkVill  -1.761398e-03
## Neighborhood_NWAmes   -5.105536e-03
## Neighborhood_NoRidge   1.345216e-02
## Neighborhood_NridgHt   1.469728e-02
## Neighborhood_OldTown  -7.050431e-03
## Neighborhood_SWISU     2.101004e-03
## Neighborhood_Sawyer   -4.217847e-03
## Neighborhood_SawyerW   3.231994e-03
## Neighborhood_Somerst   8.378889e-03
## Neighborhood_StoneBr   1.482393e-02
## Neighborhood_Timber    2.659898e-03
## Neighborhood_Veenker   4.894377e-03
## Condition1_Artery     -1.070233e-02
## Condition1_PosA       -1.267341e-03
## Condition1_PosN       -1.534387e-03
```

```
## Condition1_RRAe       -6.830604e-03
## Condition1_RRAn       -4.056844e-03
## Condition1_RRNe       -9.967417e-04
## Condition1_RRNn        2.859528e-04
## Condition2_Artery     -2.799442e-03
## Condition2_Feedr       8.057040e-04
## Condition2_PosA        1.811257e-03
## Condition2_PosN       -1.917944e-03
## BldgType_2fmCon        2.377652e-04
## BldgType_Duplex       -7.636730e-03
## BldgType_Twnhs        -9.129273e-03
## BldgType_TwnhsE       -5.600100e-03
## HouseStyle_1.5Fin      4.571342e-03
## HouseStyle_1.5Unf      2.586564e-03
## HouseStyle_2.5Unf      4.094660e-03
## HouseStyle_SFoyer     -2.741714e-04
## HouseStyle_SLvl       -1.633503e-04
## RoofStyle_Flat         6.074590e-03
## RoofStyle_Gambrel      1.402405e-03
## RoofStyle_Hip          8.174902e-04
## RoofStyle_Mansard      3.200679e-03
## RoofStyle_Shed         3.350586e-03
## RoofMatl_Tar.Grv      -3.292002e-03
## RoofMatl_WdShake       1.097280e-03
## RoofMatl_WdShngl       5.133532e-03
## Exterior1st_AsbShng   -4.350774e-05
## Exterior1st_AsphShn   -1.447028e-05
## Exterior1st_BrkComm   -7.013598e-03
## Exterior1st_BrkFace    1.005948e-02
## Exterior1st_CBlock    -2.450269e-04
## Exterior1st_CemntBd   -9.750199e-04
## Exterior1st_HdBoard   -7.903974e-03
## Exterior1st_MetalSd   -2.104612e-03
```

```
## Exterior1st_Plywood   -4.897284e-03
## Exterior1st_Stucco     1.448529e-03
## Exterior1st_Wd.Sdng   -9.878083e-03
## Exterior1st_WdShing   -3.611217e-03
## Exterior2nd_AsbShng   -3.007796e-03
## Exterior2nd_AsphShn    8.883179e-04
## Exterior2nd_Brk.Cmn   -2.048450e-03
## Exterior2nd_BrkFace   -5.437604e-03
## Exterior2nd_CBlock    -2.488318e-04
## Exterior2nd_CmentBd    1.480137e-03
## Exterior2nd_HdBoard   -6.866784e-03
## Exterior2nd_ImStucc   -7.447160e-04
## Exterior2nd_MetalSd   -2.211009e-03
## Exterior2nd_Plywood   -6.700183e-03
## Exterior2nd_Stone     -1.395584e-03
## Exterior2nd_Stucco    -9.502386e-04
## Exterior2nd_Wd.Sdng   -7.113017e-04
## Exterior2nd_Wd.Shng   -3.464635e-03
## MasVnrType_BrkCmn     -6.496428e-03
## MasVnrType_NA         -1.658933e-03
## MasVnrType_Stone       6.488697e-03
## ExterQual_Ex           2.677435e-03
## ExterQual_Fa          -1.791583e-03
## ExterCond_Ex           2.777173e-03
## ExterCond_Fa          -5.962534e-03
## ExterCond_Gd          -2.823506e-03
## ExterCond_Po          -3.071170e-03
## Foundation_BrkTil     -3.812166e-03
## Foundation_Slab       -1.548946e-03
## Foundation_Stone       4.170025e-03
## Foundation_Wood       -3.818864e-03
## BsmtQual_Ex            1.189652e-02
## BsmtQual_Fa           -1.971401e-05
```

```
## BsmtQual_NA          -7.273925e-04
## BsmtCond_Fa          -5.905748e-03
## BsmtCond_Gd           1.884160e-03
## BsmtCond_NA          -8.664239e-04
## BsmtCond_Po           1.967874e-03
## BsmtExposure_Av       5.114444e-03
## BsmtExposure_Gd       1.543709e-02
## BsmtExposure_Mn       4.348532e-03
## BsmtExposure_NA      -1.111694e-03
## BsmtFinType1_ALQ     -3.434058e-03
## BsmtFinType1_BLQ     -6.689066e-03
## BsmtFinType1_LwQ     -5.345566e-03
## BsmtFinType1_NA      -6.416287e-04
## BsmtFinType1_Unf     -4.589338e-03
## BsmtFinType2_ALQ      1.361081e-03
## BsmtFinType2_BLQ     -6.373768e-03
## BsmtFinType2_GLQ      4.861060e-03
## BsmtFinType2_NA      -8.712464e-04
## BsmtFinType2_Rec     -2.619315e-03
## Heating_GasW          5.882944e-03
## Heating_Grav         -9.186102e-03
## Heating_Wall          2.648959e-03
## HeatingQC_Fa         -2.363124e-03
## HeatingQC_Gd         -3.212118e-03
## HeatingQC_Po         -2.015421e-03
## CentralAir_N         -1.585942e-02
## Electrical_FuseA     -1.550894e-04
## Electrical_FuseF      8.786046e-04
## Electrical_FuseP     -1.747477e-03
## KitchenQual_Ex        1.694342e-02
## KitchenQual_Fa       -4.725895e-05
## Functional_Maj1      -6.325980e-03
## Functional_Maj2      -1.415851e-02
```

```
## Functional_Min1       -4.429509e-03
## Functional_Min2       -5.947916e-03
## Functional_Mod        -7.238905e-03
## Functional_Sev        -6.545929e-03
## GarageType_2Types     -5.487947e-03
## GarageType_Basment    -1.734284e-03
## GarageType_BuiltIn     1.649126e-03
## GarageType_CarPort    -1.083445e-03
## GarageType_Detchd     -8.314696e-03
## GarageType_NA         -3.570793e-03
## GarageFinish_Fin       4.899399e-03
## GarageFinish_NA       -3.639705e-03
## GarageQual_Fa         -4.037471e-03
## GarageQual_Gd          3.641942e-03
## GarageQual_NA         -3.592993e-03
## GarageQual_Po         -6.511076e-04
## GarageCond_Ex          4.562461e-04
## GarageCond_Fa         -4.876307e-03
## GarageCond_Gd         -5.952803e-04
## GarageCond_NA         -3.540699e-03
## GarageCond_Po          3.814241e-03
## PavedDrive_N          -6.086778e-03
## PavedDrive_P          -2.811773e-03
## SaleType_COD          -8.702378e-04
## SaleType_CWD           3.826170e-03
## SaleType_Con           3.426718e-03
## SaleType_ConLD         6.853327e-03
## SaleType_ConLI        -1.560323e-03
## SaleType_ConLw         2.831612e-03
## SaleType_New           8.996033e-03
## SaleType_Oth           3.198379e-03
## SaleCondition_Abnorml -1.598647e-02
## SaleCondition_AdjLand  9.154137e-04
```

```
## SaleCondition_Alloca  -1.726947e-03
## SaleCondition_Family  -6.291064e-03
## SaleCondition_Partial  5.468514e-03
## BuiltAfter1920         2.287008e-03
## YearRemodUnknown      -7.111444e-03
## NoFinBsmt             -4.724808e-03
## HasDeck                4.148937e-03
## HasPorch               8.650733e-03
```

We predict values based on our Ridge regressions.

Despite the large number of independent variables, ridge regression performs better, with a score of .14047. This puts us at 1690 out of 4216 individuals.

C. Lasso Regression To perform Lasso regression, first we define the predictor and response variables for the training dataset. Similarly to the Ridge model, we'll use the `glmnet` library, which makes it easy to use k-fold cross-validation to find the optimal value for lambda. Next, we find the coefficients for the Lasso model using our optimized lambda. Lastly, we predict new values using our optimized Lasso model. Here is our lambda:

```
## [1] 0.003096298
```

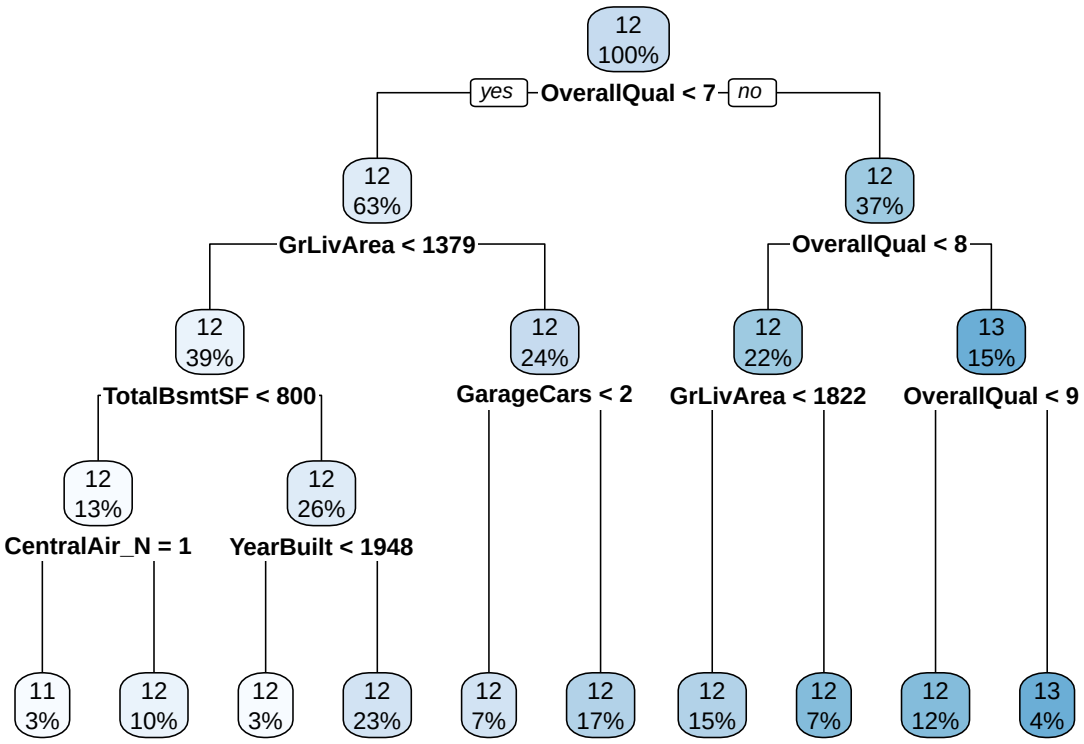
We try Lasso with both scaled and unscaled data. Because lasso incorporates a penalty based on the size of the coefficients, we expect the scaled data to perform better, and it does. Our lasso regression gives us a .1375, which outperforms ridge.

D. Elastic Net Regression In order to form elastic net, first, build a control model. Next, train the elastic net regression model. Then we optimize the elastic net model based on tuning parameters selected from model training.

Our elastic net result falls between ridge and lasso.

E. Basic Decision Tree After Elastic Net we tried a basic Decision Tree model. It scored 0.22422 so clearly not as good of a model as those previously used, including our base model.

```
##
## Regression tree:
## rpart(formula = SalePrice ~ ., data = dfTrain6)
##
## Variables actually used in tree construction:
## [1] CentralAir_N GarageCars   GrLivArea    OverallQual  TotalBsmtSF
## [6] YearBuilt
##
## Root node error: 228.26/1456 = 0.15677
##
## n= 1456
##
##          CP nsplit rel error  xerror     xstd
## 1  0.463479      0   1.00000 1.00071 0.042774
## 2  0.078148      1   0.53652 0.53795 0.025866
## 3  0.075204      2   0.45837 0.48684 0.023782
## 4  0.045004      3   0.38317 0.39150 0.020042
## 5  0.021131      4   0.33816 0.35264 0.016416
## 6  0.018292      5   0.31703 0.32707 0.015952
## 7  0.015909      6   0.29874 0.31861 0.015540
## 8  0.015487      7   0.28283 0.30445 0.013940
## 9  0.012442      8   0.26735 0.29386 0.013800
## 10 0.010000      9   0.25490 0.28752 0.013534
```



F. Other tree-based models: Random Forest and Gradient Boosting Our final models are Random Forest and Gradient Boosting, which also make use of decision trees.

Below are the top variables for our Gradient Boosting model:

	var	rel.inf
OverallQual	OverallQual	34.3575946
GrLivArea	GrLivArea	16.2567195
TotalBsmtSF	TotalBsmtSF	6.5914685
YearBuilt	YearBuilt	5.0145990
GarageArea	GarageArea	3.1105262
YearRemodAdd	YearRemodAdd	2.9637372
GarageCars	GarageCars	2.9187880
LotArea	LotArea	2.8725270
X1stFlrSF	X1stFlrSF	2.8370188
BsmtFinSF1	BsmtFinSF1	2.5497933

	var	rel.inf
OverallCond	OverallCond	2.1754019
Fireplaces	Fireplaces	1.6886810
CentralAir_N	CentralAir_N	1.5705128
BsmtUnfSF	BsmtUnfSF	1.0441738
Id	Id	0.9348620
GarageYrBlt	GarageYrBlt	0.8106067
X2ndFlrSF	X2ndFlrSF	0.7253192
OpenPorchSF	OpenPorchSF	0.6028966
MSZoning_RM	MSZoning_RM	0.5644134
SaleCondition_Abnorml	SaleCondition_Abnorml	0.5478962

Below is the output from our random forest model:

```
##
## Call:
##  randomForest(formula = SalePrice ~ ., data = dfTrain6)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 72
##
##           Mean of squared residuals: 0.01754065
##                     % Var explained: 88.81
```

	IncNodePurity
OverallQual	58.898232
GrLivArea	34.985490
YearBuilt	19.916129
TotalBsmtSF	11.876984
GarageCars	11.096525
GarageArea	10.717971

	IncNodePurity
X1stFlrSF	8.849367
FullBath	7.801985
GarageYrBlt	7.016787
LotArea	4.669041
YearRemodAdd	4.397052
Fireplaces	4.358778
BsmtFinSF1	4.053526
X2ndFlrSF	3.248898
CentralAir_N	2.501134
OverallCond	2.392884
BsmtQual_Ex	1.774957
TotRmsAbvGrd	1.712547
BsmtUnfSF	1.372916
OpenPorchSF	1.324895

The results of our Random Forest model looks very similar to that of our Gradient Boosting, but the model does not improve our score, while Gradient Boosting does. With Gradient Boosting we land on .12786, which puts us in the 80th percentile.

Discussion and Conclusions

Ordinary Least Squares is a regression technique with a long history of use as a predictive model. However, standard measures of fit (like R^2) will always increase (or stay the same) as you add independent variables. This can result in models which incorporate noise - in other words, overfit the data so that idiosyncrasies in the training set affect predictions in the test set. Other methods of measuring fit, such as adjusted R^2 and AIC, help mitigate the overfitting effect by penalizing the addition of factors.

More recently, other techniques which employ regularization have been introduced to deal with overfit. For example, in ridge regression, we reduce the sum of our coefficients, not the number of variables. We do this by introducing a penalty in the loss function represented by the squared sum of the coefficients themselves, multiplied by a factor (designated as lambda) which allows us to control the degree to which the size of the coefficients matters. If lambda is zero, there is no difference between ridge regression and OLS.

Ridge regression will keep all the variables but may significantly reduce the coefficients for some. Lasso regression is similar in that it employs a constraint where the sum of the absolute value of the coefficients is less than a fixed value. Lasso regression may drop coefficients altogether to stay under the constraint.

Elastic Net regression is a hybrid approach that blends both of the penalizations of lasso and ridge methods. An alpha parameter weights which penalty to emphasize - lasso or ridge.

Decision trees, including Random Forest and Gradient Boosting as discussed below, incorporate sequential choice-point steps, providing different outcomes for each choice-point. Decision trees have the advantage that they do not make assumptions that the dependent variable is linearly related to the independent variables. They can also be graphically represented, as has been done with the Basic Decision Tree above, to help users more easily interpret the model and understand the basic decisions made as part of the supervised learning process.

Random Forest (RF) and Gradient Boosting (GB) both combine multiple trees so the results are averages of many samples, which improves their predictability. However, the results may be difficult to interpret. RF and GB handle the combination of trees differently. RF builds each tree independently and averages at the end. BG proceeds in a stage-wise manner, improving the performance of weak learners as you go. This can result in better performance.

RF and BG handle the combination of trees differently. RF builds each tree independently and averages at the end. BG proceeds in a stage-wise manner, improving the performance of weak learners as you go. This can result in better performance.

Our dataset has features that lend to overfitting. Most significant of these is the high number of potential independent variables (over 200 once the dummy variables are created.) Multicollinearity is also a problem, though less than we might have expected.

We used stepAIC to fit our OLS model. StepAIC uses backward substitution to find the best model with the lowest AIC. With an adjusted R^2 of over 90% overfitting was expected. However, even with an overfit model our predictions performed at the 60th percentile on the Kaggle.

Because of the large number of potential predictors, ridge (and by extension elastic net) were not as good candidates as Lasso - however, potential issues with collinearity actually favored ridge. We found that Lasso improved our score the most of the regression-based models, followed be elastic net (which is a compromise between lasso and ridge), followed by ridge. All were improvements over OLS - however, the improvements were not dramatic.

Gradient boosting had the most success. Because gradient boosting is a machine-learning technique in which the model receives direct feedback with each iteration, it can often do a better job of predicting than the other models. However, if we were looking for insight into the data, GB is something of a "black box" which makes interpretation difficult.

Our gradient boosting model relied on a few key variables - overall quality, size, number of cars and year built. The fact that number of cars and size of garage both featured prominently suggests we could have improved the model by eliminating some multicollinearity (the two are highly correlated).

In conclusion, it is important to keep in mind that while regularization improved our model, the base OLS model also performed adequately, so regularization, while important, may in some cases improve models at the margin. It is also important to recognize the strengths of each of the techniques and use the appropriate one for the situation.

References

Alfiyatin, A. N. (2017, December 1). *Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study : Malang, East Java, Indonesia.* Https://Thesai.Org/. https://thesai.org/Publications/ViewPaper?Volume=8&Issue=10&Code=IJACSA&SerialNo=42

Guan, J. (2021, November 12). *Predicting home sale prices: A review of existing methods and illustration of data stream methods for improved performance.* University of Louisville College of Business. https://business.louisville.edu/faculty-research/research-publications/predicting-home-sale-prices-a-review-of-existing-methods-and-illustration-of-data-stream-methods-for-improved-performance/

Journal, I. (2019, May 4). *Predicting housing prices using advanced regression techniques.* Ijariit Journal - Academia.Edu. https://www.academia.edu/39014594/Predicting_housing_prices_using_advanced_regression_techniques#:%7E:text=There%20are%20various%20techniques%20for%20predicting%20house%20prices.,have%20an%20impact%20on%20a%20topic%20of%

Kennedy, J. (2014, June 11). *Particle swarm optimization.* Https://Www.Academia.Edu. https://www.academia.edu/1446115/Particle_swarm_optimization

Li, D. (2021, July 3). *Prediction of China's Housing Price Based on a Novel Grey Seasonal Model.* Www.Hindawi.Com. https://www.hindawi.com/journals/mpe/2021/5541233/ Liu, S. (2011, September 1). A brief introduction to Grey systems theory. Https://Www.Researchgate.Net.

https://www.researchgate.net/publication/252052256_A_brief_introduction_to_Grey_systems_theory

Liu, X. (2012, January 14). *Spatial and Temporal Dependence in House Price Prediction.* Springer-Link. https://link.springer.com/article/10.1007/s11146-011-9359-3?error=cookies_not_supported&code=d2a7946f-1472-4dd7-9b57-50d3eba69e24

Wu, Z., et. al., (2020, November 5). *Prediction of California House Price Based on Multiple Linear Regression* | Francis Academic Press. Https://Www.Academia.Edu/. https://francis-press.com/papers/2868