621_Final_HomeSales

Eric Hirsch, Cameron Smith and Carlisle Fergusen

4/7/2022

Contents

Abstract 2

Introduction 2

Background and Literature Review 2

Modeling 3

1. Dataset Description 4

    A. Summary Statistics 4

    B. Missing values 10

    C. Create dummy variables 12

    D. Reconcile training and test sets 12

    E. Multicollinearity 12

2. Transformations 13

    A. Log of SalePrice 13

    B. Other transformations 13

3. Model and Predict: 14

    A. Base Model 14

    B. Now we try Ridge regression: 15

C. Lasso Regression 23

D. Elastic Net Regression 23

Discussion and Conclusions 24

References 25

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 4 standard regression techniques (OLS, ridge lasso, and elastic net) 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 Lasso to be the best predictor, which we speculate is because the dataset has a high number of predictors relative to the number of observations.

Introduction

In this paper we analyze housing prices by comparing three prediction methodologies: OLS, Ridge regression, and Random Forest. 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. [1] Predictive models 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]

[1] 1 Li, 2021
[2] 2 Journal, 2019

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 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 three types of regression: OLS, Ridge, and Random Forest. At 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 discussed 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 presense of a pool, etc. The response variable, "SalePrice", is a continuous variable representing the sale price of the house in dollars.

We examine the data:

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

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
##
```

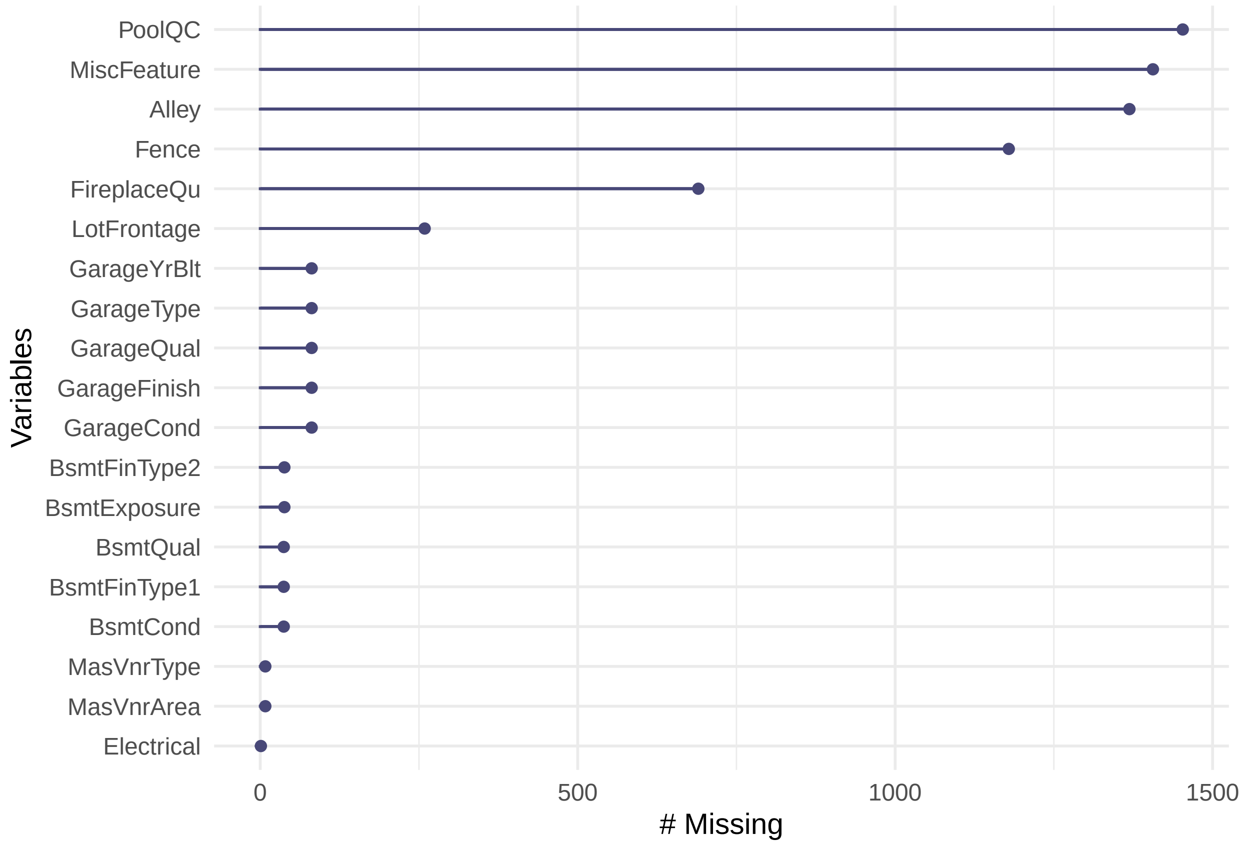
```
## 'data.frame':    1460 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50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

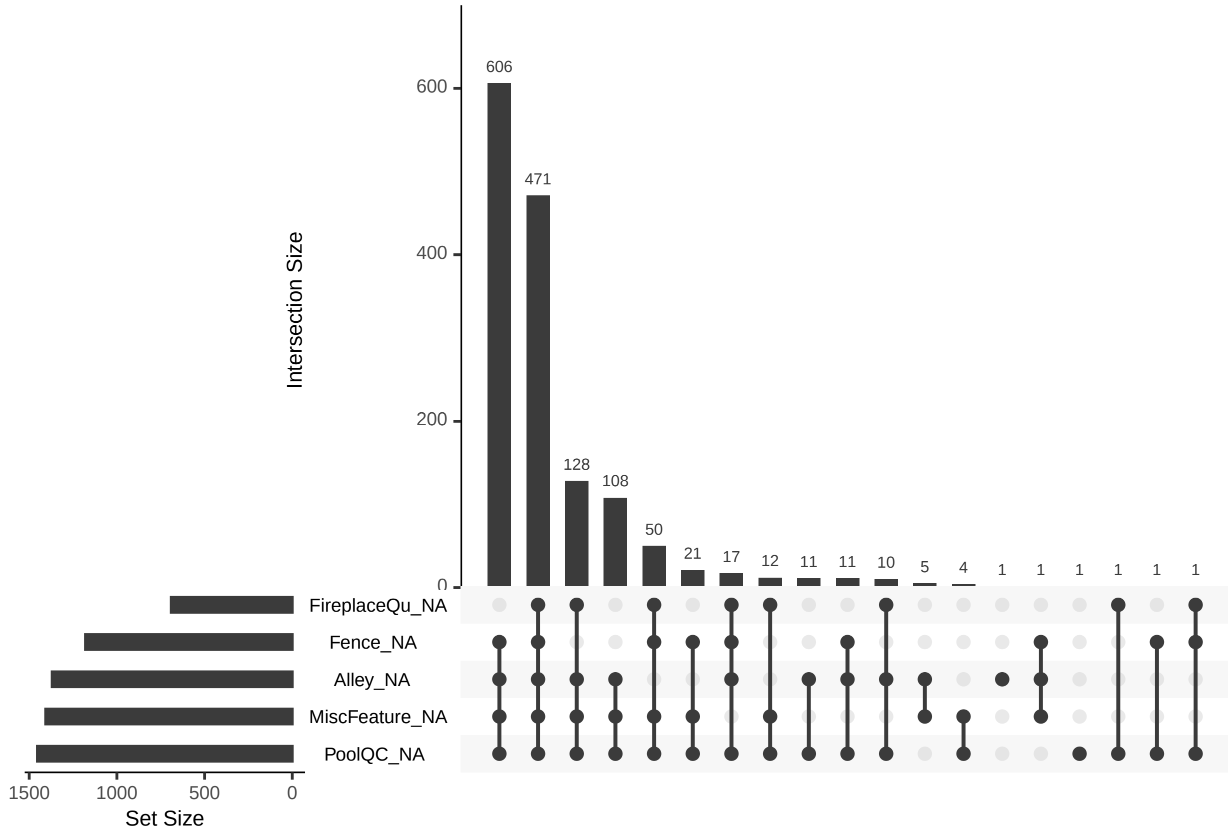
```
## $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...
## $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...
## $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...
## $ MasVnrArea   : int  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:


PoolQC
MiscFeature
Alley
Fence
FireplaceQu
LotFrontage
GarageYrBlt
GarageType
GarageQual
GarageFinish
GarageCond
BsmtFinType2
BsmtExposure
BsmtQual
BsmtFinType1
BsmtCond
MasVnrType
MasVnrArea
Electrical
Variables
0
500
1000
1500
Missing



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
```

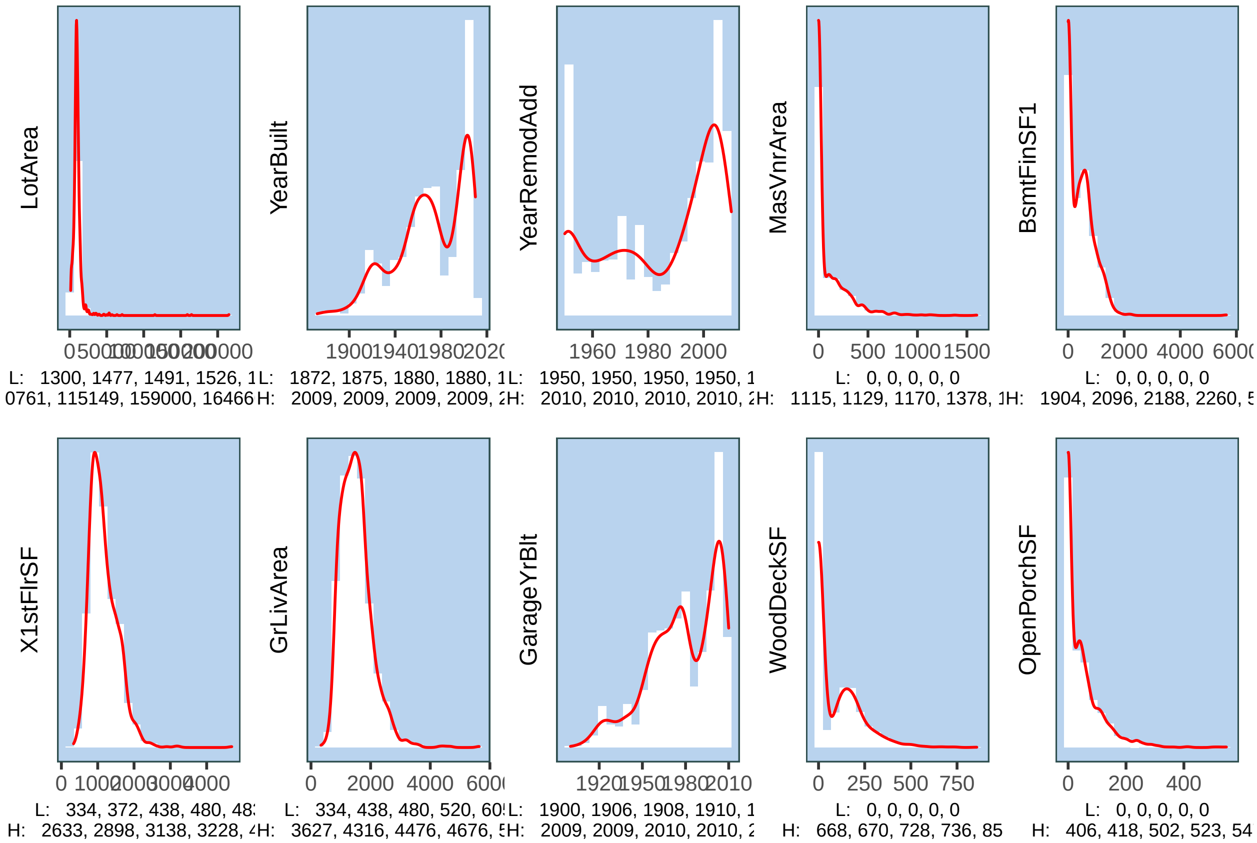
```
## 3            GrLivArea          TotRmsAbvGrd   0.8254894
## 5           GarageCars            GarageArea   0.8824754
## 7          MSZoning_FV  Neighborhood_Somerst   0.8628071
## 9       RoofStyle_Flat      RoofMatl_Tar.Grv   0.8349139
## 11 Exterior1st_AsbShng   Exterior2nd_AsbShng   0.8479167
## 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 a regression using the stepAIC algorithm to minimize AIC.

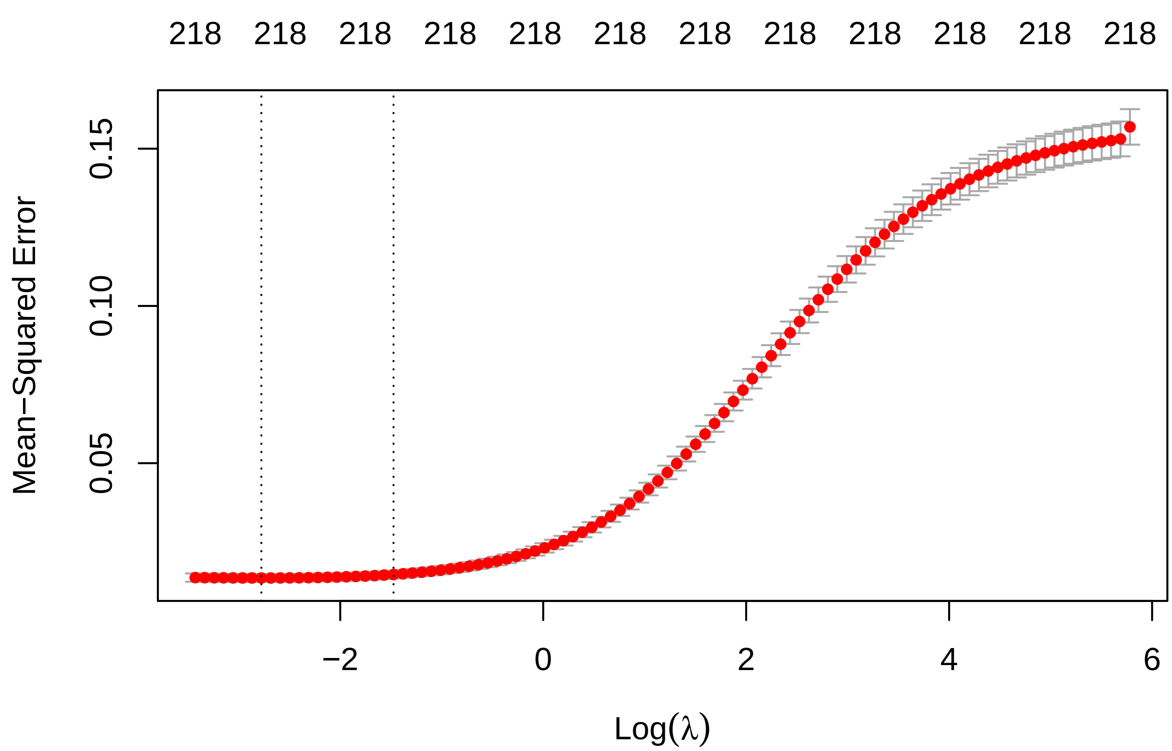
```
##
## Call:
## lm(formula = SalePrice ~ GrLivArea, data = dfTrain6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31671 -0.14546  0.03394  0.16266  0.90722
```

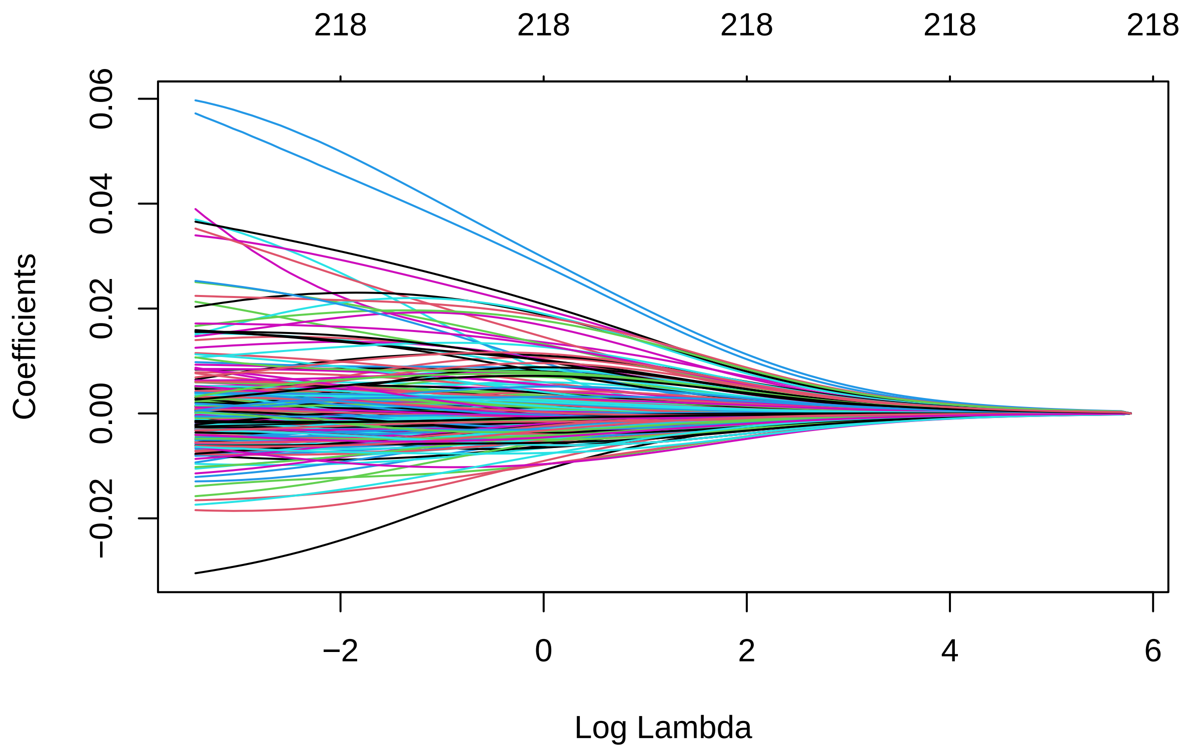
```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.116e+01  2.306e-02  483.99   <2e-16 ***
## GrLivArea   5.731e-04  1.453e-05   39.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2754 on 1454 degrees of freedom
## Multiple R-squared:  0.5167, Adjusted R-squared:  0.5164
## F-statistic:  1555 on 1 and 1454 DF,  p-value: < 2.2e-16
```

Now we make predictions

We achieve a score of .14586 on kaggle.

B. Now we try Ridge regression: R makes it easy to find the best lambda by using kfold validation:






```
## 219 x 1 sparse Matrix of class "dgCMatrix"
##                                 s0
## (Intercept)           1.202194e+01
## Id                   -3.316167e-03
## MSSubClass           -1.220442e-03
## LotArea               1.893276e-02
## OverallQual           5.585056e-02
## OverallCond           3.307967e-02
## YearBuilt             2.921600e-02
## YearRemodAdd          8.396676e-03
## MasVnrArea            4.463805e-03
## BsmtFinSF1            2.321538e-02
## BsmtFinSF2            3.394083e-03
## BsmtUnfSF             5.868867e-03
## TotalBsmtSF           3.205511e-02
```

```
## X1stFlrSF              3.398655e-02
## X2ndFlrSF              3.074557e-02
## LowQualFinSF           1.455037e-03
## GrLivArea              5.219902e-02
## BsmtFullBath           1.503678e-02
## BsmtHalfBath           8.263591e-04
## FullBath               2.215637e-02
## HalfBath               1.461155e-02
## BedroomAbvGr           1.862441e-03
## KitchenAbvGr          -1.237620e-02
## TotRmsAbvGrd           1.832210e-02
## Fireplaces             1.635576e-02
## GarageYrBlt            8.430868e-03
## GarageCars             2.208106e-02
## GarageArea             1.816561e-02
## WoodDeckSF             9.495562e-03
## OpenPorchSF            5.407171e-03
## EnclosedPorch          5.396358e-03
## X3SsnPorch             5.918845e-03
## ScreenPorch            1.087578e-02
## PoolArea               5.224207e-03
## MiscVal               -1.893104e-03
## MoSold                -7.361292e-04
## YrSold                -1.847161e-03
## MSZoning_C..all.      -2.818916e-02
## MSZoning_FV            7.785702e-03
## MSZoning_RM           -1.298121e-02
## Street_Grvl           -5.527020e-03
## LotShape_IR1           1.066411e-03
## LotShape_IR2           4.638174e-03
## LotShape_IR3           1.459866e-03
## LandContour_Bnk       -1.353820e-03
## LandContour_HLS        2.920167e-03
```

```
## LandContour_Low       -1.090528e-03
## LotConfig_Corner       2.845655e-03
## LotConfig_CulDSac      9.062037e-03
## LotConfig_FR2         -5.179079e-03
## LotConfig_FR3         -1.621573e-03
## LandSlope_Mod          2.208949e-03
## LandSlope_Sev         -7.297180e-03
## Neighborhood_Blmngtn  -6.953252e-05
## Neighborhood_Blueste  -2.935210e-03
## Neighborhood_BrDale   -8.574068e-03
## Neighborhood_BrkSide   6.155806e-03
## Neighborhood_ClearCr   5.698727e-03
## Neighborhood_Crawfor   2.352916e-02
## Neighborhood_Edwards  -9.885964e-03
## Neighborhood_Gilbert   7.866620e-04
## Neighborhood_IDOTRR   -2.564449e-03
## Neighborhood_MeadowV  -1.860140e-02
## Neighborhood_Mitchel  -4.492634e-03
## Neighborhood_NPkVill  -1.720593e-03
## Neighborhood_NWAmes   -5.253035e-03
## Neighborhood_NoRidge   1.338503e-02
## Neighborhood_NridgHt   1.505907e-02
## Neighborhood_OldTown  -6.906287e-03
## Neighborhood_SWISU     2.222073e-03
## Neighborhood_Sawyer   -3.956776e-03
## Neighborhood_SawyerW   3.338802e-03
## Neighborhood_Somerst   8.494840e-03
## Neighborhood_StoneBr   1.509195e-02
## Neighborhood_Timber    2.700942e-03
## Neighborhood_Veenker   4.898939e-03
## Condition1_Artery     -1.112093e-02
## Condition1_PosA       -1.514734e-03
## Condition1_PosN       -1.596212e-03
```

```
## Condition1_RRAe       -7.061452e-03
## Condition1_RRAn       -4.276013e-03
## Condition1_RRNe       -1.093182e-03
## Condition1_RRNn        2.407893e-04
## Condition2_Artery     -2.773141e-03
## Condition2_Feedr       9.605498e-04
## Condition2_PosA        1.834850e-03
## Condition2_PosN       -2.011688e-03
## BldgType_2fmCon        2.964283e-04
## BldgType_Duplex       -7.780375e-03
## BldgType_Twnhs        -9.402093e-03
## BldgType_TwnhsE       -5.771063e-03
## HouseStyle_1.5Fin      4.654782e-03
## HouseStyle_1.5Unf      2.673297e-03
## HouseStyle_2.5Unf      4.151987e-03
## HouseStyle_SFoyer     -1.464736e-04
## HouseStyle_SLvl        4.438741e-05
## RoofStyle_Flat         6.634384e-03
## RoofStyle_Gambrel      1.404284e-03
## RoofStyle_Hip          6.641289e-04
## RoofStyle_Mansard      3.335277e-03
## RoofStyle_Shed         3.566185e-03
## RoofMatl_Tar.Grv      -3.690499e-03
## RoofMatl_WdShake       9.022830e-04
## RoofMatl_WdShngl       5.057080e-03
## Exterior1st_AsbShng    2.512051e-04
## Exterior1st_AsphShn   -4.090004e-05
## Exterior1st_BrkComm   -7.193092e-03
## Exterior1st_BrkFace    1.033384e-02
## Exterior1st_CBlock    -2.707114e-04
## Exterior1st_CemntBd   -1.295719e-03
## Exterior1st_HdBoard   -7.948707e-03
## Exterior1st_MetalSd   -2.068717e-03
```

```
## Exterior1st_Plywood   -4.989657e-03
## Exterior1st_Stucco     1.456822e-03
## Exterior1st_Wd.Sdng   -1.036136e-02
## Exterior1st_WdShing   -3.589031e-03
## Exterior2nd_AsbShng   -3.280628e-03
## Exterior2nd_AsphShn    8.795836e-04
## Exterior2nd_Brk.Cmn   -1.936721e-03
## Exterior2nd_BrkFace   -5.905631e-03
## Exterior2nd_CBlock    -2.688550e-04
## Exterior2nd_CmentBd    1.735964e-03
## Exterior2nd_HdBoard   -7.017346e-03
## Exterior2nd_ImStucc   -7.981273e-04
## Exterior2nd_MetalSd   -2.238328e-03
## Exterior2nd_Plywood   -6.822890e-03
## Exterior2nd_Stone     -1.609856e-03
## Exterior2nd_Stucco    -1.137528e-03
## Exterior2nd_Wd.Sdng   -5.344159e-04
## Exterior2nd_Wd.Shng   -3.570795e-03
## MasVnrType_BrkCmn     -6.458243e-03
## MasVnrType_NA         -1.812043e-03
## MasVnrType_Stone       6.463089e-03
## ExterQual_Ex           2.235792e-03
## ExterQual_Fa          -1.532623e-03
## ExterCond_Ex           2.744351e-03
## ExterCond_Fa          -5.985140e-03
## ExterCond_Gd          -3.107162e-03
## ExterCond_Po          -2.937623e-03
## Foundation_BrkTil     -3.913867e-03
## Foundation_Slab       -1.827664e-03
## Foundation_Stone       4.366354e-03
## Foundation_Wood       -4.061401e-03
## BsmtQual_Ex            1.168843e-02
## BsmtQual_Fa            1.667580e-04
```

```
## BsmtQual_NA           -8.707477e-04
## BsmtCond_Fa           -5.883071e-03
## BsmtCond_Gd            1.737232e-03
## BsmtCond_NA           -5.556071e-04
## BsmtCond_Po            2.419696e-03
## BsmtExposure_Av        5.234627e-03
## BsmtExposure_Gd        1.558444e-02
## BsmtExposure_Mn        4.357142e-03
## BsmtExposure_NA       -7.403299e-04
## BsmtFinType1_ALQ      -3.363694e-03
## BsmtFinType1_BLQ      -6.627262e-03
## BsmtFinType1_LwQ      -5.481796e-03
## BsmtFinType1_NA       -3.027530e-04
## BsmtFinType1_Unf      -4.652304e-03
## BsmtFinType2_ALQ       1.262671e-03
## BsmtFinType2_BLQ      -6.504581e-03
## BsmtFinType2_GLQ       4.993026e-03
## BsmtFinType2_NA       -1.054292e-03
## BsmtFinType2_Rec      -2.637787e-03
## Heating_GasW           6.000639e-03
## Heating_Grav          -9.456538e-03
## Heating_Wall           2.946812e-03
## HeatingQC_Fa          -2.119679e-03
## HeatingQC_Gd          -3.263399e-03
## HeatingQC_Po          -2.113756e-03
## CentralAir_N          -1.613808e-02
## Electrical_FuseA       1.827542e-04
## Electrical_FuseF       1.142244e-03
## Electrical_FuseP      -1.829331e-03
## KitchenQual_Ex         1.705759e-02
## KitchenQual_Fa         1.002915e-04
## Functional_Maj1       -6.519887e-03
## Functional_Maj2       -1.458709e-02
```

```
## Functional_Min1       -4.510770e-03
## Functional_Min2       -6.020738e-03
## Functional_Mod        -7.597360e-03
## Functional_Sev        -6.813091e-03
## GarageType_2Types     -5.564597e-03
## GarageType_Basment    -1.708063e-03
## GarageType_BuiltIn     1.362803e-03
## GarageType_CarPort    -8.057846e-04
## GarageType_Detchd     -8.014494e-03
## GarageType_NA         -3.602695e-03
## GarageFinish_Fin       4.571995e-03
## GarageFinish_NA       -3.459337e-03
## GarageQual_Fa         -4.233673e-03
## GarageQual_Gd          3.634157e-03
## GarageQual_NA         -3.349563e-03
## GarageQual_Po         -9.464443e-04
## GarageCond_Ex          5.371144e-04
## GarageCond_Fa         -4.888778e-03
## GarageCond_Gd         -6.181713e-04
## GarageCond_NA         -3.375997e-03
## GarageCond_Po          4.154394e-03
## PavedDrive_N          -6.128714e-03
## PavedDrive_P          -2.954823e-03
## SaleType_COD          -5.850902e-04
## SaleType_CWD           3.857856e-03
## SaleType_Con           3.469501e-03
## SaleType_ConLD         7.285072e-03
## SaleType_ConLI        -1.536375e-03
## SaleType_ConLw         2.940252e-03
## SaleType_New           9.316847e-03
## SaleType_Oth           3.430097e-03
## SaleCondition_Abnorml -1.638142e-02
## SaleCondition_AdjLand  9.840009e-04
```

```
## SaleCondition_Alloca  -1.718252e-03
## SaleCondition_Family  -6.308048e-03
## SaleCondition_Partial  5.022071e-03
## BuiltAfter1920         2.075557e-03
## YearRemodUnknown      -6.790909e-03
## NoFinBsmt             -4.420746e-03
## HasDeck                3.695499e-03
## HasPorch               8.224311e-03
```

We predict values based on our Ridge regressions.

Ridge regression performs the best, 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.

```
## [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.

```
## gbm(formula = SalePrice ~ ., distribution = "gaussian", data = dfTrain6,
##     n.trees = 10000, interaction.depth = 4, shrinkage = 0.01)
## A gradient boosted model with gaussian loss function.
## 10000 iterations were performed.
## There were 218 predictors of which 144 had non-zero influence.
```

```
##
## 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: 0.01756745
##                     % Var explained: 88.79
```

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.

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, 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.

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.* SpringerLink. 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