Housing Prices Regression

Zhiwei Lin

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(glmnet)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-6
```

Import data

```
train<- read_csv("train.csv")
```

```
## Rows: 1460 Columns: 81
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConf...
## dbl (38): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
test <- read_csv("test.csv")
```

```
## Rows: 1459 Columns: 80
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConf...
## dbl (37): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
head(train)
```

```
## # A tibble: 6 x 81
##      Id MSSubClass MSZoning LotFr~1 LotArea Street Alley LotSh~2 LandC~3 Utili~4
##   <dbl>      <dbl> <chr>      <dbl>   <dbl> <chr>  <chr> <chr>   <chr>   <chr>
## 1     1         60 RL            65    8450 Pave   <NA>  Reg     Lvl     AllPub
## 2     2         20 RL            80    9600 Pave   <NA>  Reg     Lvl     AllPub
## 3     3         60 RL            68   11250 Pave   <NA>  IR1     Lvl     AllPub
## 4     4         70 RL            60    9550 Pave   <NA>  IR1     Lvl     AllPub
## 5     5         60 RL            84   14260 Pave   <NA>  IR1     Lvl     AllPub
## 6     6         50 RL            85   14115 Pave   <NA>  IR1     Lvl     AllPub
## # ... with 71 more variables: LotConfig <chr>, LandSlope <chr>,
## #   Neighborhood <chr>, Condition1 <chr>, Condition2 <chr>, BldgType <chr>,
## #   HouseStyle <chr>, OverallQual <dbl>, OverallCond <dbl>, YearBuilt <dbl>,
## #   YearRemodAdd <dbl>, RoofStyle <chr>, RoofMatl <chr>, Exterior1st <chr>,
## #   Exterior2nd <chr>, MasVnrType <chr>, MasVnrArea <dbl>, ExterQual <chr>,
## #   ExterCond <chr>, Foundation <chr>, BsmtQual <chr>, BsmtCond <chr>,
## #   BsmtExposure <chr>, BsmtFinType1 <chr>, BsmtFinSF1 <dbl>, ...
```

```
test$Id <- NULL
train$Id <- NULL
test$SalePrice <- NA
```

Remove id variable in both data and add SalePrice variable to test_data

```
all <- rbind(train,test)
```

combine train and test data

```
missing_percentage <- function(df){
 colSums(is.na(df))/nrow(df)
}
missing_percentage(all)
```

```
##    MSSubClass      MSZoning   LotFrontage       LotArea        Street 
##  0.0000000000  0.0013703323  0.1664953751  0.0000000000  0.0000000000 
##         Alley      LotShape   LandContour     Utilities     LotConfig 
##  0.9321685509  0.0000000000  0.0000000000  0.0006851662  0.0000000000 
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType 
##  0.0000000000  0.0000000000  0.0000000000  0.0000000000  0.0000000000 
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd 
##  0.0000000000  0.0000000000  0.0000000000  0.0000000000  0.0000000000 
##     RoofStyle      RoofMatl   Exterior1st   Exterior2nd    MasVnrType 
##  0.0000000000  0.0000000000  0.0003425831  0.0003425831  0.0082219938 
##    MasVnrArea     ExterQual     ExterCond    Foundation      BsmtQual 
##  0.0078794108  0.0000000000  0.0000000000  0.0000000000  0.0277492292 
##      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2 
##  0.0280918123  0.0280918123  0.0270640630  0.0003425831  0.0274066461 
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC 
##  0.0003425831  0.0003425831  0.0003425831  0.0000000000  0.0000000000 
##    CentralAir    Electrical      1stFlrSF      2ndFlrSF  LowQualFinSF 
##  0.0000000000  0.0003425831  0.0000000000  0.0000000000  0.0000000000 
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath 
##  0.0000000000  0.0006851662  0.0006851662  0.0000000000  0.0000000000 
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional 
##  0.0000000000  0.0000000000  0.0003425831  0.0000000000  0.0006851662 
##    Fireplaces   FireplaceQu    GarageType   GarageYrBlt  GarageFinish 
##  0.0000000000  0.4864679685  0.0537855430  0.0544707091  0.0544707091 
##    GarageCars    GarageArea    GarageQual    GarageCond    PavedDrive 
##  0.0003425831  0.0003425831  0.0544707091  0.0544707091  0.0000000000 
##    WoodDeckSF   OpenPorchSF EnclosedPorch     3SsnPorch   ScreenPorch 
##  0.0000000000  0.0000000000  0.0000000000  0.0000000000  0.0000000000 
##      PoolArea        PoolQC         Fence   MiscFeature       MiscVal 
##  0.0000000000  0.9965741692  0.8043850634  0.9640287770  0.0000000000 
##        MoSold        YrSold      SaleType SaleCondition     SalePrice 
##  0.0000000000  0.0000000000  0.0003425831  0.0000000000  0.4998287085
```

```r
all <- mutate_if(all,is.character,as.factor)
```

```r
all <- all %>% mutate_if(is.factor, ~ ifelse(is.na(.), 0, .))
# replace missing values NA with 0 for all categorical variables
all <- all %>% mutate_if(is.numeric, ~ ifelse(is.na(.), mean(., na.rm = TRUE), .))
# replcae missing values NA with mean for all numeric variables
sum(is.na(all))
```

```
## [1] 0
```

```r
# no missing values in the data anymore
```

```r
# Split the data back to training and test sets
train_data <- all[1:nrow(train),]
test_data <- all[(nrow(train)+1):nrow(all),]
test_data$SalePrice <- NA
```

Ridge regression

```
lambda <- 10^seq(-3, 3, length = 100)
```

```
# Build the model
set.seed(123)
ridge <- train(
  SalePrice ~., data = train_data, method = "glmnet",
  trControl = trainControl("cv", number = 10),
  tuneGrid = expand.grid(alpha = 0, lambda = lambda)
  )
```

```
# Model coefficients
coef(ridge$finalModel, ridge$bestTune$lambda)
```

```
## 80 x 1 sparse Matrix of class "dgCMatrix"
##                           s1
## (Intercept)    1.590079e+06
## MSSubClass    -7.574294e+01
## MSZoning      -1.844319e+03
## LotFrontage   -9.841805e+01
## LotArea        3.734902e-01
## Street         3.032689e+04
## Alley         -2.813798e+03
## LotShape      -9.757778e+02
## LandContour    2.073182e+03
## Utilities     -4.185069e+04
## LotConfig     -7.806494e+01
## LandSlope      4.862401e+03
## Neighborhood   2.660318e+02
## Condition1    -5.690022e+02
## Condition2    -8.223009e+03
## BldgType      -2.344855e+03
## HouseStyle    -6.583748e+02
## OverallQual    1.003744e+04
## OverallCond    4.432523e+03
## YearBuilt      1.357571e+02
## YearRemodAdd   7.911962e+01
## RoofStyle      1.852354e+03
## RoofMatl       4.723247e+03
## Exterior1st   -6.856764e+02
## Exterior2nd    2.502328e+02
## MasVnrType     3.615820e+03
## MasVnrArea     3.005154e+01
## ExterQual     -9.857949e+03
## ExterCond      6.589677e+02
## Foundation     8.568175e+02
## BsmtQual      -6.822881e+03
## BsmtCond       2.850036e+03
## BsmtExposure  -2.723548e+03
## BsmtFinType1  -4.268888e+02
## BsmtFinSF1     9.808956e+00
## BsmtFinType2   1.846611e+03
## BsmtFinSF2     1.331640e+01
## BsmtUnfSF      6.243585e-01
## TotalBsmtSF    1.299931e+01
## Heating       -1.747944e+03
## HeatingQC     -7.918285e+02
## CentralAir     2.534156e+03
## Electrical    -4.588454e+02
## ‘1stFlrSF‘     1.577211e+01
## ‘2ndFlrSF‘     1.807983e+01
## LowQualFinSF  -1.495945e+01
## GrLivArea      2.073197e+01
## BsmtFullBath   6.348293e+03
## BsmtHalfBath   3.517593e+02
## FullBath       4.834663e+03
## HalfBath       1.821559e+03
## BedroomAbvGr  -2.706703e+03
## KitchenAbvGr  -1.633139e+04
## KitchenQual   -8.110780e+03
## TotRmsAbvGrd   3.658318e+03
## Functional     3.531496e+03
## Fireplaces     7.304832e+03
## FireplaceQu   -7.982269e+02
## GarageType     1.609407e+02
## GarageYrBlt   -2.854904e+01
## GarageFinish  -2.211134e+03
## GarageCars     9.092250e+03
## GarageArea     1.238980e+01
## GarageQual    -1.298970e+03
## GarageCond    -1.108356e+02
## PavedDrive     2.735088e+03
## WoodDeckSF     2.080849e+01
## OpenPorchSF   -4.903567e+00
## EnclosedPorch  3.774804e+00
## ‘3SsnPorch‘    2.456111e+01
## ScreenPorch    4.672345e+01
## PoolArea       2.883107e+02
## PoolQC        -9.073622e+04
## Fence          3.012748e+02
## MiscFeature   -1.666935e+03
## MiscVal        2.634597e-01
## MoSold        -9.728437e+01
## YrSold        -9.670621e+02
## SaleType      -5.210729e+02
## SaleCondition  2.832771e+03
```

```
# Make predictions
ridge_predictions <- ridge %>% predict(test_data)
ridge_predictions <- list(unname(ridge_predictions))[[1]]
head(ridge_predictions)
```

```
## [1] 112625.4 162275.5 172171.1 189024.2 190119.3 173952.5
```

Lasso regression

```
# Build the model
set.seed(123)
lasso <- train(
  SalePrice ~., data = train_data, method = "glmnet",
  trControl = trainControl("cv", number = 10),
  tuneGrid = expand.grid(alpha = 1, lambda = lambda)
  )
```

```
# Model coefficients
coef(lasso$finalModel, lasso$bestTune$lambda)
```

```
## 80 x 1 sparse Matrix of class "dgCMatrix"
##                           s1
## (Intercept)    9.977898e+05
## MSSubClass    -7.683223e+01
## MSZoning      -2.043587e+03
## LotFrontage   -1.080784e+02
## LotArea        4.071036e-01
## Street         3.054043e+04
## Alley         -1.952668e+03
## LotShape      -6.947967e+02
## LandContour    1.580685e+03
## Utilities     -3.728149e+04
## LotConfig      .
## LandSlope      3.831658e+03
## Neighborhood   2.109665e+02
## Condition1    -4.698519e+02
## Condition2    -8.296996e+03
## BldgType      -2.555701e+03
## HouseStyle    -3.878252e+02
## OverallQual    1.072505e+04
## OverallCond    5.444117e+03
## YearBuilt      2.011788e+02
## YearRemodAdd   1.530859e+01
## RoofStyle      9.606591e+02
## RoofMatl       4.221183e+03
## Exterior1st   -7.280742e+02
## Exterior2nd    3.150269e+02
## MasVnrType     3.974552e+03
## MasVnrArea     3.139741e+01
## ExterQual     -1.004066e+04
## ExterCond      4.923942e+02
## Foundation     .
## BsmtQual      -7.120689e+03
## BsmtCond       2.546859e+03
## BsmtExposure  -2.567291e+03
## BsmtFinType1  -2.161994e+02
## BsmtFinSF1     1.049722e+01
## BsmtFinType2   1.749219e+03
## BsmtFinSF2     1.215882e+01
## BsmtUnfSF      .
## TotalBsmtSF    1.438338e+01
## Heating       -1.108243e+03
## HeatingQC     -4.386006e+02
## CentralAir     1.097620e+03
## Electrical    -2.356310e+02
## ‘1stFlrSF‘     .
## ‘2ndFlrSF‘     1.103718e+00
## LowQualFinSF  -4.180105e+01
## GrLivArea      4.215983e+01
## BsmtFullBath   5.654138e+03
## BsmtHalfBath  -2.169250e+02
## FullBath       1.283185e+03
## HalfBath       .
## BedroomAbvGr  -3.631023e+03
## KitchenAbvGr  -1.630775e+04
## KitchenQual   -7.622710e+03
## TotRmsAbvGrd   3.796544e+03
## Functional     3.474986e+03
## Fireplaces     6.945320e+03
## FireplaceQu   -9.284502e+02
## GarageType     5.402206e+01
## GarageYrBlt    .
## GarageFinish  -1.841193e+03
## GarageCars     1.005255e+04
## GarageArea     4.374141e+00
## GarageQual    -1.467178e+03
## GarageCond     .
## PavedDrive     2.033219e+03
## WoodDeckSF     2.009377e+01
## OpenPorchSF   -2.662077e+00
## EnclosedPorch -9.375567e-01
## ‘3SsnPorch‘    2.176285e+01
## ScreenPorch    4.588821e+01
## PoolArea       6.689632e+02
## PoolQC        -1.894544e+05
## Fence          .
## MiscFeature   -1.247885e+03
## MiscVal        .
## MoSold        -3.487299e+01
## YrSold        -7.049526e+02
## SaleType      -4.622954e+02
## SaleCondition  3.383327e+03
```

```
# Make predictions
lasso_predictions <- lasso %>% predict(test_data)
lasso_predictions <- list(unname(lasso_predictions))[[1]]
head(lasso_predictions)
```

```
## [1] 113048.1 160854.5 169713.7 187489.0 189715.1 172390.3
```

elastic net regession

```
# Build the model using the training set
set.seed(123)
elastic <- train(
  SalePrice ~., data = train_data, method = "glmnet",
  trControl = trainControl("cv", number = 10),
  tuneLength = 10
  )
```

```
# Model coefficients
coef(elastic$finalModel, elastic$bestTune$lambda)
```

```
## 80 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)   -2.222545e+04
## MSSubClass    -6.017465e+01
## MSZoning      -1.147227e+02
## LotFrontage    .
## LotArea        3.086747e-01
## Street         1.485696e+04
## Alley         -1.669626e+02
## LotShape      -8.007844e+02
## LandContour    5.167040e+02
## Utilities     -1.200510e+02
## LotConfig      .
## LandSlope      1.805793e+03
## Neighborhood   1.167375e+02
## Condition1     .
## Condition2    -3.122362e+03
## BldgType      -1.504731e+03
## HouseStyle     .
## OverallQual    1.083895e+04
## OverallCond    2.791447e+03
## YearBuilt      1.040328e+02
## YearRemodAdd   1.092165e+02
## RoofStyle      1.181253e+03
## RoofMatl       3.916179e+03
## Exterior1st   -1.144936e+02
## Exterior2nd    .
## MasVnrType     2.009820e+03
## MasVnrArea     2.466299e+01
## ExterQual     -9.872008e+03
## ExterCond      .
## Foundation     .
## BsmtQual      -5.811176e+03
## BsmtCond       2.137000e+03
## BsmtExposure  -1.667971e+03
## BsmtFinType1   .
## BsmtFinSF1     9.999052e+00
## BsmtFinType2   .
## BsmtFinSF2     8.251030e-02
## BsmtUnfSF      .
## TotalBsmtSF    1.419577e+01
## Heating        .           
## HeatingQC     -6.978261e+02
## CentralAir     2.879013e+03
## Electrical     .           
## `1stFlrSF`     1.222391e+01
## `2ndFlrSF`     1.121039e+01
## LowQualFinSF   .           
## GrLivArea      2.288009e+01
## BsmtFullBath   5.441548e+03
## BsmtHalfBath   .           
## FullBath       3.709265e+03
## HalfBath       1.472197e+03
## BedroomAbvGr   .           
## KitchenAbvGr  -1.312546e+04
## KitchenQual   -8.632634e+03
## TotRmsAbvGrd   2.473421e+03
## Functional     2.736324e+03
## Fireplaces     5.938355e+03
## FireplaceQu    .           
## GarageType     .           
## GarageYrBlt    .           
## GarageFinish  -1.747602e+03
## GarageCars     8.321729e+03
## GarageArea     1.196946e+01
## GarageQual     .           
## GarageCond     .           
## PavedDrive     1.646628e+03
## WoodDeckSF     1.938248e+01
## OpenPorchSF    .           
## EnclosedPorch  .           
## `3SsnPorch`    .           
## ScreenPorch    3.195121e+01
## PoolArea       7.667022e+01
## PoolQC        -3.737457e+04
## Fence          .           
## MiscFeature   -1.342096e+01
## MiscVal        .           
## MoSold         .           
## YrSold        -1.992799e+02
## SaleType       .           
## SaleCondition  1.765328e+03
```

```
# Make predictions
elastic_predictions <- elastic %>% predict(test_data)
elastic_predictions <- list(unname(elastic_predictions))[[1]]
head(elastic_predictions)
```

```
## [1] 113815.0 164734.5 175006.6 192678.4 192390.1 175869.5
```

Comparing models

```
models <- list(ridge = ridge, lasso = lasso, elastic = elastic)
resamples(models) %>% summary( metric = "RMSE")
```

```
##
## Call:
## summary.resamples(object = ., metric = "RMSE")
##
## Models: ridge, lasso, elastic
## Number of resamples: 10
##
## RMSE
##             Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## ridge   25043.56 26561.72 31352.83 33808.79 36573.45 61047.36    0
## lasso   24407.04 26443.11 32237.38 34461.08 37888.28 60703.50    0
## elastic 25732.45 26415.56 30501.21 33821.52 36788.75 61461.48    0
```

```
submission<-read_csv("sample_submission.csv")
```

```
## Rows: 1459 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Id, SalePrice
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
submission<-mutate(submission, SalePrice=elastic_predictions)
write.csv(submission, file = "submission.csv",row.names = FALSE)
```