NYC Data Science Academy

George Goginashvili

Project 3

The goal of the project is to assist investment decision making in identifying and selecting parts (independent variables) of a real property that contribute to its value the most. In order to achieve this, I have used linear models, regularized linear models, and tree based models. I have also transformed and structured data in order to better apply the models and improve their validity.

Data Learning and Transformation

My first step in learning of data was to investigate qualitative part of each independent variable to better understand their possible statistical relationship with property values. This step has also shown that there was a possibility that these variables were correlated among each other which could create multicollinearity problems in developing predictive models. Since the variables included numeric and character variables, I have also checked if the values of the variables included NAs. The data did indeed included NA. Consequently, based on my investigation of the data, I have developed the rules to assign values to NAs: if a variable was numeric, I have assigned zero or average value, and if a variable was character, I have assigned ‘None”, “No such value” like "NoAlley", or average level like “Mix”. Finally, I have converted all character variables into factors.

Data Analysis

Before I have dived into developing the models, I wanted to have a general sense of relationship of independent variables with property values. I ran a linear model where I included all independent variables. At this step, I haven’t split data into train and test parts. The model has shown Adjusted R-squared of 92% which

confirmed statistical relationship of independent variables with the dependent variable. However, the model has shown that a big part of independent variables were not statistically significant, presumably because of their correlation among each other which supported my assumptions of multicollinearity problems. In order to screen variables at this step, I have run a linear model for each independent variable and dependent variable, and utilized only those variables that had Adjusted R-squared higher than 10%.

The screening gave 35 independent variables which I have included into my linear model:

```
model.Model.First = lm(SalePrice ~ MSSubClass+MSZoning+Neighborhood+OverallQual+
                                   OverallCond+YearBuilt+YearRemodAdd+Exterior1st+Exterior2nd+
                                   MasVnrType+MasVnrArea+ExterQual+Foundation+BsmtQual+
                                   BsmtExposure+BsmtFinType1+BsmtFinSF1+TotalBsmtSF+HeatingQC+
                                   X1stFlrSF+X2ndFlrSF+GrLivArea+FullBath+KitchenQual+
                                   TotRmsAbvGrd+Fireplaces+FireplaceQu+GarageType+GarageYrBlt+
                                   GarageFinish+GarageCars+GarageArea+WoodDeckSF+SaleType+
                                    SaleCondition, data = train)
```

The model has given Adjusted R-squared of 87%. However, the model has also shown that some of the variables still were not statistically significant. In order to check multicollinearity, I have used following commands:

- cor(numeric.variables)
- plot(numeric.variables)
- pairs.panels(numeric.variables, color="red")
- ggpairs(numeric.variables)

all of them confirmed that some numeric variables had high correlation among each other.

I have further screened the variables using stepwise selections for P-values and AIC.

Stepwise selection for P-values selected the following variables:

OverallQual, GrLivArea, Neighborhood, BsmtExposure, MSSubClass, KitchenQual, GarageCars, OverallCond, YearBuilt, BsmtFinType1, BsmtQual, Fireplaces, SaleCondition, FullBath, Exterior1st, WoodDeckSF, X2ndFlrSF

Stepwise selection for AIC selected the following variables:

OverallQual, GrLivArea, Neighborhood, MSSubClass, BsmtExposure, GarageCars, OverallCond, BsmtFinType1, BsmtQual, SaleCondition, Fireplaces, YearBuilt, FullBath, Exterior1st, WoodDeckSF, X2ndFlrSF, MSZoning, MasVnrArea

Running linear models has given Adjusted R-squared of 87% each. Each model also has shown that excluding the rest of the variables had no statistical impact on outcomes. However, some variables in each model still were not partially significant (some levels of factor variables were not significant).

I have further screened above variables based on statistical significance and qualitative assumptions not to have multicollinearity. I have received the following two models:

First Model

```
Call:
lm(formula = SalePrice ~ GrLivArea + GarageCars + YearBuilt +
    BsmtQual + Fireplaces + WoodDeckSF + X2ndFlrSF, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-457655  -18291    -749   17007  260572

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.129e+05  1.077e+05  -7.547 7.85e-14 ***
GrLivArea     7.956e+01  3.413e+00  23.313  < 2e-16 ***
GarageCars    1.677e+04  1.824e+03   9.195  < 2e-16 ***
YearBuilt     4.609e+02  5.409e+01   8.520  < 2e-16 ***
BsmtQualFa   -8.135e+04  8.721e+03  -9.328  < 2e-16 ***
BsmtQualGd   -6.646e+04  4.126e+03 -16.107  < 2e-16 ***
BsmtQualNone -1.093e+05  7.918e+03 -13.801  < 2e-16 ***
BsmtQualTA   -7.811e+04  5.077e+03 -15.385  < 2e-16 ***
Fireplaces    1.252e+04  1.811e+03   6.911 7.17e-12 ***
WoodDeckSF    3.029e+01  8.547e+00   3.544 0.000407 ***
X2ndFlrSF    -2.237e+01  3.409e+00  -6.562 7.36e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38520 on 1449 degrees of freedom
Multiple R-squared:  0.7665,    Adjusted R-squared:  0.7649
F-statistic: 475.7 on 10 and 1449 DF,  p-value: < 2.2e-16
```

Second Model

```
Call:
lm(formula = SalePrice ~ GrLivArea + KitchenQual + GarageCars + 
    BsmtQual + Fireplaces + YearBuilt + WoodDeckSF + X2ndFlrSF + 
    MSZoning + MasVnrArea, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-445045  -16772    -929   15462  258512 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -3.028e+05  1.118e+05  -2.709 0.006826 ** 
GrLivArea      6.541e+01  3.355e+00  19.494  < 2e-16 ***
KitchenQualFa -6.515e+04  7.706e+03  -8.455  < 2e-16 ***
KitchenQualGd -4.381e+04  4.499e+03  -9.737  < 2e-16 ***
KitchenQualTA -6.366e+04  4.824e+03 -13.196  < 2e-16 ***
GarageCars     1.456e+04  1.729e+03   8.418  < 2e-16 ***
BsmtQualFa    -6.537e+04  8.275e+03  -7.899 5.53e-15 ***
BsmtQualGd    -4.698e+04  4.284e+03 -10.968  < 2e-16 ***
BsmtQualNone  -8.674e+04  7.682e+03 -11.292  < 2e-16 ***
BsmtQualTA    -5.794e+04  5.105e+03 -11.350  < 2e-16 ***
Fireplaces     1.043e+04  1.708e+03   6.104 1.33e-09 ***
YearBuilt      2.109e+02  5.620e+01   3.752 0.000182 ***
WoodDeckSF     2.916e+01  8.011e+00   3.640 0.000282 ***
X2ndFlrSF     -1.622e+01  3.264e+00  -4.968 7.56e-07 ***
MSZoningFV     4.334e+04  1.258e+04   3.445 0.000587 ***
MSZoningRH     3.084e+04  1.453e+04   2.123 0.033944 *  
MSZoningRL     3.959e+04  1.161e+04   3.411 0.000666 ***
MSZoningRM     2.547e+04  1.166e+04   2.184 0.029094 *  
MasVnrArea     3.085e+01  6.033e+00   5.113 3.60e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35790 on 1441 degrees of freedom
Multiple R-squared:  0.7996,    Adjusted R-squared:  0.7971 
F-statistic: 319.4 on 18 and 1441 DF,  p-value: < 2.2e-16
```

Test of validity of both models has shown that because of a few outliers normality and homoscedasticity were violated but acceptable (knowing the shortcoming of the model because of outliers).

Having conducted this preliminary analysis and acquiring knowledge how independent variables are correlated, I have moved forward to further develop the models.

Linear Models

For linear models, I have split the data on train data consisting 70% of the original data and test data consisting the rest 30% of the data. I have followed the same steps for linear models as above (screening using only Adjusted R-squired higher than 10% and stepwise selection for P-values and AIC), but only with train data. It gave me same independent variables for each of the two linear models as above in complete data.

First Model for P-values

```
Call:
lm(formula = SalePrice ~ GrLivArea + GarageCars + YearBuilt +
    KitchenQual + BsmtQual + Fireplaces + WoodDeckSF + X2ndFlrSF,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-411069  -17550    -216   14536  266080

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.042e+05  1.242e+05  -4.863 1.34e-06 ***
GrLivArea      6.244e+01  3.856e+00  16.192  < 2e-16 ***
GarageCars     1.705e+04  2.049e+03   8.323 2.76e-16 ***
YearBuilt      3.786e+02  6.205e+01   6.102 1.49e-09 ***
KitchenQualFa -6.658e+04  9.407e+03  -7.078 2.74e-12 ***
KitchenQualGd -4.220e+04  5.511e+03  -7.659 4.39e-14 ***
KitchenQualTA -6.194e+04  5.901e+03 -10.496  < 2e-16 ***
BsmtQualFa    -5.521e+04  9.587e+03  -5.759 1.12e-08 ***
BsmtQualGd    -4.028e+04  5.187e+03  -7.767 1.98e-14 ***
BsmtQualNone  -6.908e+04  9.693e+03  -7.127 1.96e-12 ***
BsmtQualTA    -4.801e+04  6.177e+03  -7.772 1.90e-14 ***
Fireplaces     1.259e+04  2.033e+03   6.195 8.46e-10 ***
WoodDeckSF     2.884e+01  9.304e+00   3.100  0.00199 **
X2ndFlrSF     -1.616e+01  3.805e+00  -4.247 2.37e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36390 on 1008 degrees of freedom
Multiple R-squared:  0.7776,    Adjusted R-squared:  0.7747
F-statistic: 271.1 on 13 and 1008 DF,  p-value: < 2.2e-16
```

Second Model for AIC

```
Call:
lm(formula = SalePrice ~ GrLivArea + KitchenQual + GarageCars +
    BsmtQual + Fireplaces + YearBuilt + WoodDeckSF + X2ndFlrSF +
    MSZoning + MasVnrArea, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-404215  -16340    -814   14923  275361

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.807e+05  1.350e+05  -2.820 0.004899 **
GrLivArea      5.888e+01  3.948e+00  14.911  < 2e-16 ***
KitchenQualFa -6.480e+04  9.418e+03  -6.880 1.05e-11 ***
KitchenQualGd -4.244e+04  5.455e+03  -7.780 1.80e-14 ***
KitchenQualTA -6.251e+04  5.841e+03 -10.701  < 2e-16 ***
GarageCars     1.727e+04  2.043e+03   8.454  < 2e-16 ***
BsmtQualFa    -5.807e+04  9.496e+03  -6.115 1.38e-09 ***
BsmtQualGd    -4.103e+04  5.154e+03  -7.961 4.59e-15 ***
BsmtQualNone  -7.460e+04  9.668e+03  -7.717 2.88e-14 ***
BsmtQualTA    -5.072e+04  6.149e+03  -8.248 5.00e-16 ***
Fireplaces     1.130e+04  2.038e+03   5.543 3.81e-08 ***
YearBuilt      2.417e+02  6.728e+01   3.593 0.000343 ***
WoodDeckSF     2.779e+01  9.268e+00   2.998 0.002784 **
X2ndFlrSF     -1.416e+01  3.855e+00  -3.674 0.000252 ***
MSZoningFV     5.969e+04  1.942e+04   3.074 0.002168 **
MSZoningRH     4.404e+04  2.192e+04   2.010 0.044751 *
MSZoningRL     5.440e+04  1.852e+04   2.937 0.003386 **
MSZoningRM     3.941e+04  1.861e+04   2.117 0.034479 *
MasVnrArea     1.509e+01  7.231e+00   2.087 0.037131 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35970 on 1003 degrees of freedom
Multiple R-squared:  0.7838,    Adjusted R-squared:  0.7799
F-statistic:   202 on 18 and 1003 DF,  p-value: < 2.2e-16
```

The models also repeated lack validity regarding normality and homoscedasticity. Consequently, I have decided to drop outliers to adjust validity of the models to an acceptable level.

After dropping outliers, I have gotten the following models with significant increases of values of Adjusted R-squared for each.

Model for P-values

```
Call:
lm(formula = SalePrice ~ GrLivArea + GarageCars + YearBuilt +
    KitchenQual + BsmtQual + Fireplaces + WoodDeckSF + X2ndFlrSF,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-107427  -16923    -140   14768  178617

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -7.539e+05  1.063e+05  -7.090 2.52e-12 ***
GrLivArea      8.203e+01  3.546e+00  23.133  < 2e-16 ***
GarageCars     1.262e+04  1.771e+03   7.126 1.97e-12 ***
YearBuilt      4.470e+02  5.307e+01   8.423  < 2e-16 ***
KitchenQualFa -6.716e+04  8.017e+03  -8.377  < 2e-16 ***
KitchenQualGd -4.422e+04  4.709e+03  -9.390  < 2e-16 ***
KitchenQualTA -6.139e+04  5.031e+03 -12.203  < 2e-16 ***
BsmtQualFa    -5.071e+04  8.171e+03  -6.207 7.91e-10 ***
BsmtQualGd    -4.128e+04  4.425e+03  -9.329  < 2e-16 ***
BsmtQualNone  -6.833e+04  8.258e+03  -8.274 4.08e-16 ***
BsmtQualTA    -4.761e+04  5.264e+03  -9.044  < 2e-16 ***
Fireplaces     1.063e+04  1.739e+03   6.110 1.42e-09 ***
WoodDeckSF     2.454e+01  7.935e+00   3.093  0.00204 **
X2ndFlrSF     -2.795e+01  3.312e+00  -8.439  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31000 on 1005 degrees of freedom
Multiple R-squared:   0.83,     Adjusted R-squared:  0.8278
F-statistic: 377.5 on 13 and 1005 DF,  p-value: < 2.2e-16
```

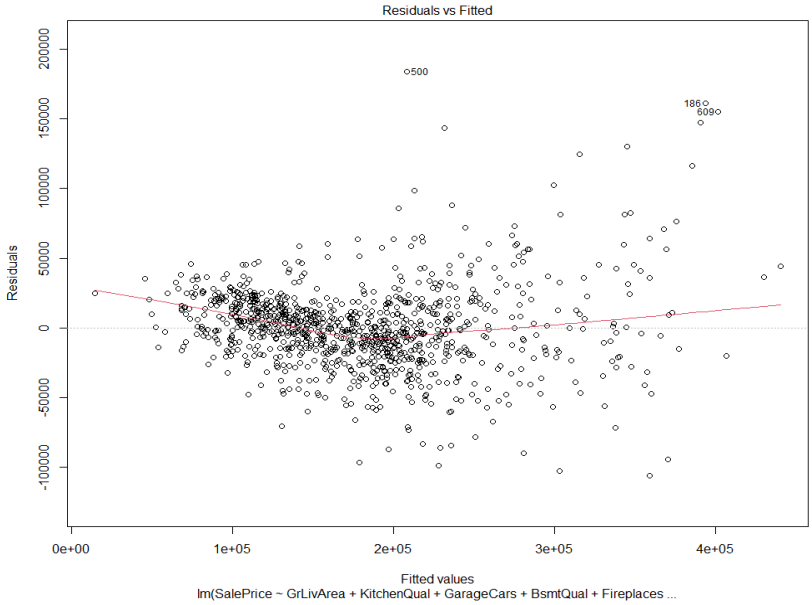
Model for AIC

```
Call:
lm(formula = SalePrice ~ GrLivArea + KitchenQual + GarageCars +
    BsmtQual + Fireplaces + YearBuilt + WoodDeckSF + X2ndFlrSF +
    MSZoning + MasVnrArea, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-105694  -16156    -479   15053  183739

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.254e+05  1.149e+05  -4.574 5.40e-06 ***
GrLivArea      7.783e+01  3.603e+00  21.600  < 2e-16 ***
KitchenQualFa -6.502e+04  7.972e+03  -8.156 1.03e-15 ***
KitchenQualGd -4.389e+04  4.631e+03  -9.477  < 2e-16 ***
KitchenQualTA -6.150e+04  4.947e+03 -12.432  < 2e-16 ***
GarageCars     1.249e+04  1.756e+03   7.113 2.17e-12 ***
BsmtQualFa    -5.321e+04  8.040e+03  -6.618 5.93e-11 ***
BsmtQualGd    -4.102e+04  4.368e+03  -9.391  < 2e-16 ***
BsmtQualNone  -7.250e+04  8.182e+03  -8.861  < 2e-16 ***
BsmtQualTA    -4.962e+04  5.204e+03  -9.535  < 2e-16 ***
Fireplaces     9.412e+03  1.732e+03   5.435 6.87e-08 ***
YearBuilt      3.090e+02  5.723e+01   5.399 8.38e-08 ***
WoodDeckSF     2.402e+01  7.849e+00   3.060 0.002274 **
X2ndFlrSF     -2.644e+01  3.338e+00  -7.922 6.19e-15 ***
MSZoningFV     5.796e+04  1.643e+04   3.527 0.000439 ***
MSZoningRH     4.027e+04  1.855e+04   2.171 0.030158 *
MSZoningRL     5.030e+04  1.567e+04   3.209 0.001375 **
MSZoningRM     3.860e+04  1.575e+04   2.451 0.014419 *
MasVnrArea     2.581e+01  6.159e+00   4.191 3.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30440 on 1000 degrees of freedom
Multiple R-squared:  0.837,     Adjusted R-squared:  0.8341
F-statistic: 285.3 on 18 and 1000 DF,  p-value: < 2.2e-16
```

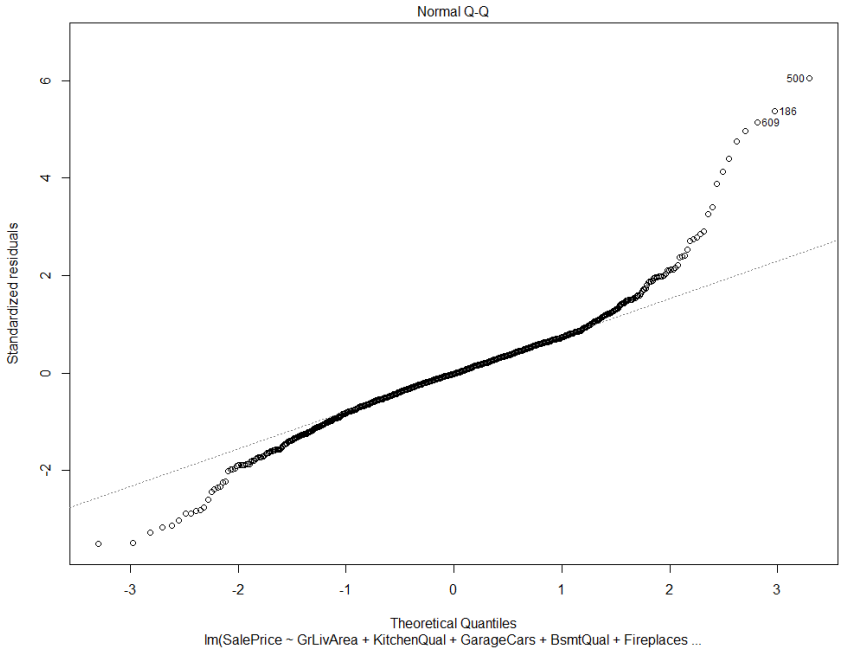
From above two models, I have chosen Model for AIC with Adjusted R-squared of 83%.

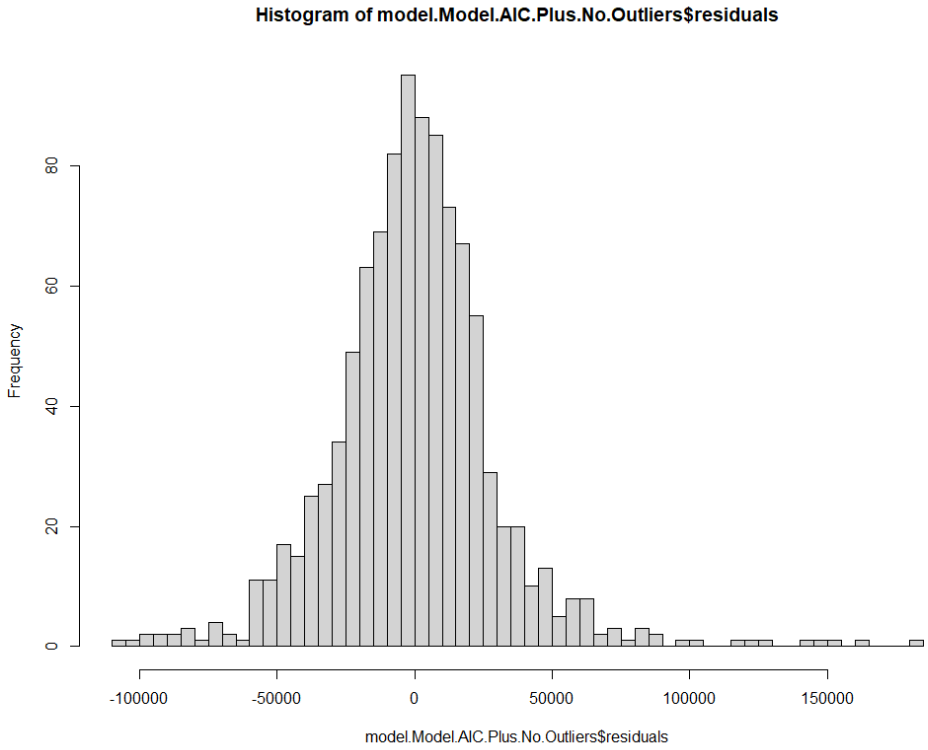
Regarding validity assumptions, I have gotten the following:

- Equality of variances is a bit violated but acceptable

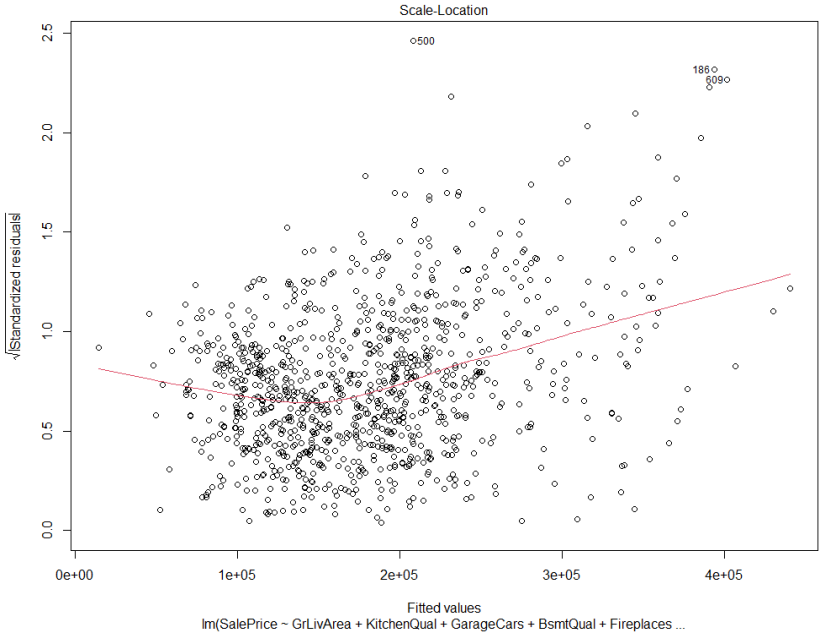


- Normality is also violated a bit. However the following picture shows that it is acceptable



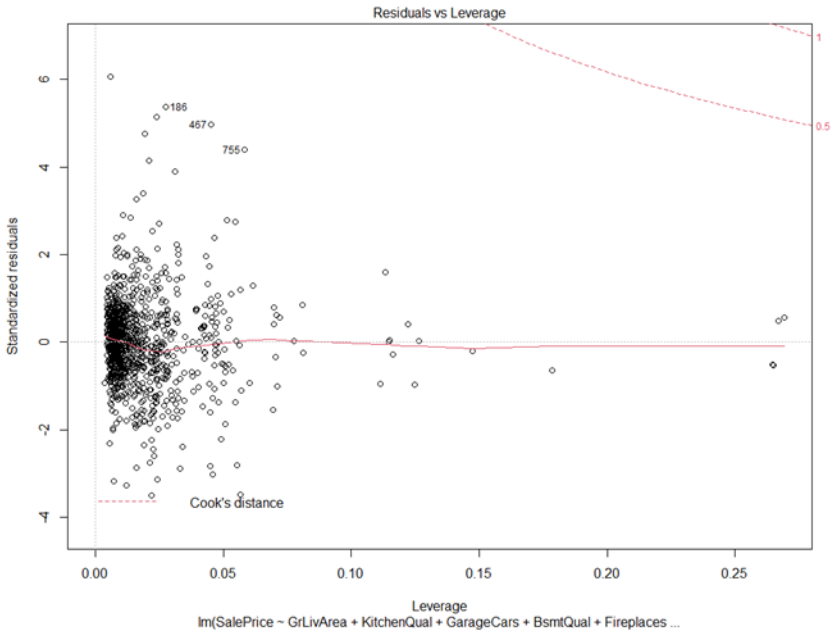


- The picture below shows that homoscedasticity is a bit violated but acceptable because VIF of each variable is acceptable



	GVIF	Df	GVIF^(1/(2*Df))
GrLivArea	3.519851	1	1.876127
KitchenQual	2.492286	3	1.164393
GarageCars	1.921008	1	1.386004
BsmtQual	3.808103	4	1.181921
Fireplaces	1.366491	1	1.168970
YearBuilt	3.239418	1	1.799838
WoodDeckSF	1.145238	1	1.070158
X2ndFlrSF	2.293686	1	1.514492
MSZoning	1.616696	4	1.061888
MasVnrArea	1.290842	1	1.136152

- The model continues to have outliers, but they don't have significant impact on its validity. See below picture



The final part of chosen linear model is correlation of predicted values from test data to its actual values which is 91.69%. I believe such correlation shows a good quality of the final chosen predictive linear model.

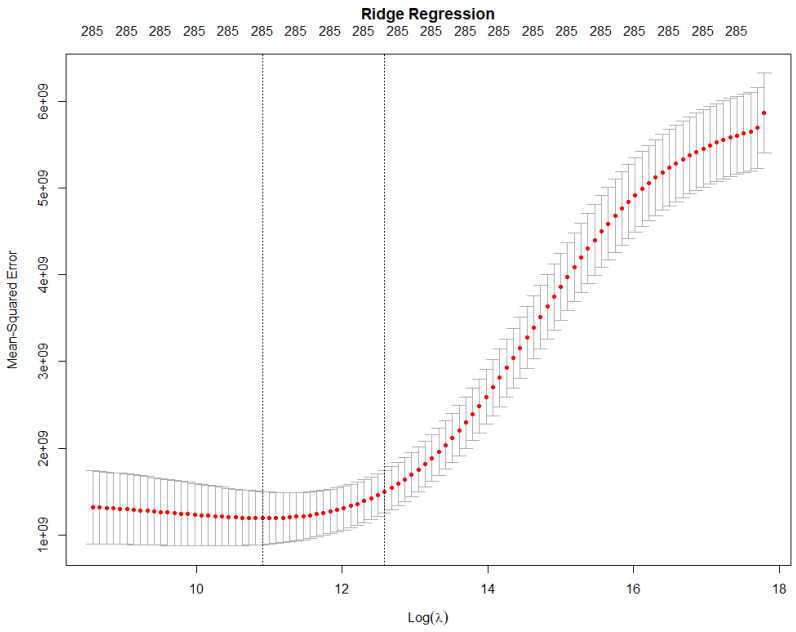
Based on the final linear model, we have the following independent variables having significant statistical relationship with the dependent variable – property value:

- GrLivArea: Above grade (ground) living area square feet
- KitchenQual: Kitchen quality
- GarageCars: Size of garage in car capacity
- BsmtQual: Evaluates the height of the basement
- Fireplaces: Number of fireplaces
- YearBuilt: Original construction date
- WoodDeckSF: Wood deck area in square feet
- X2ndFlrSF: Second floor square feet
- MSZoning: Identifies the general zoning classification of the sale
- MasVnrArea: Masonry veneer area in square feet

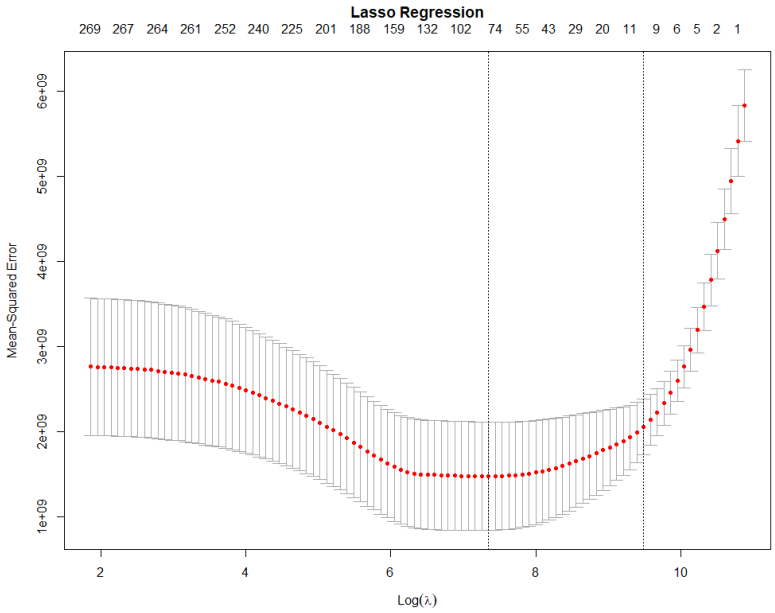
Linear Regularized Models

I have utilized Ridge Regression, Lasso Regression, and Elastic Net.

Cross validation at Ridge Regression has shown that the best log lambda is at 12.58069. Please see picture below. At best alpha, correlation of predicted values from test data and actual values is 90.65% which is lower than correlation of 91.69 achieved in linear model for AIC.



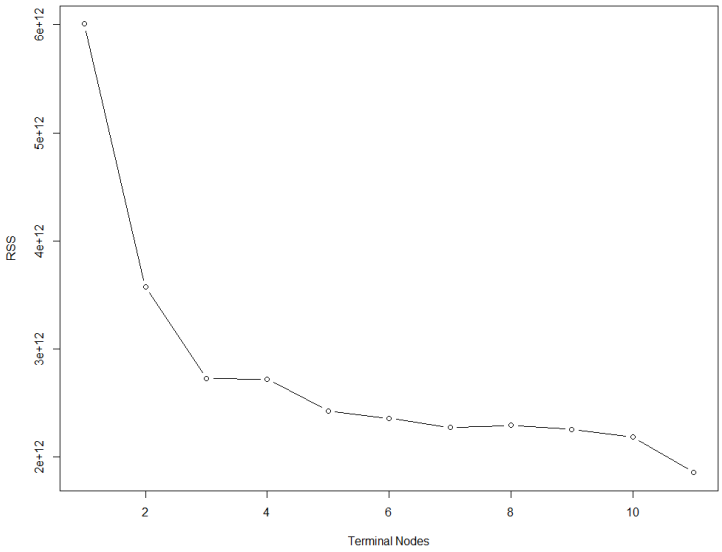
Cross validation for Lasso Regression has provided best value for log lambda of 9.487314. See picture below. Correlation achieved at best log alpha with test data equals 89.36%.



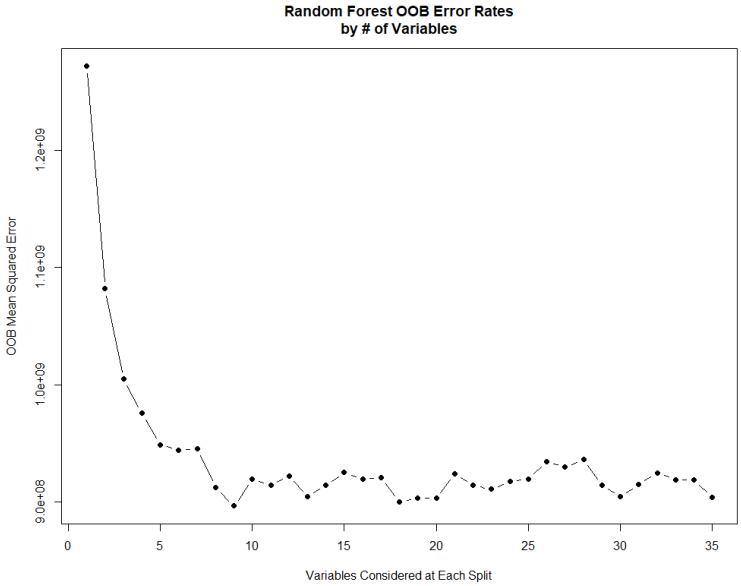
Regularized linear model – Elastic Net achieves highest correlation of 91.59% between predicted values from test data and actual values from test data at alpha = 0.1 which is still lower than correlation of 91.69% achieved at multivariate linear model.

Tree Based Models

Cross validation of tree models has shown that first time significant reductions in RSS the models achieved at 7 terminal nodes. See picture below. However, the lowest RSS has been achieved at 11 terminal nodes. Consequently, in order to check which model has highest correlation with test data, I have pruned the tree at 7 terminal nodes, and used also default 11 terminal nodes. At 7 terminal nodes, the tree model achieved the correlation of 83.09% between predicted values and actual values from test data. At default 11 nodes, the tree model achieved the correlation of 87.73%.



The analysis of random forests using different number of variables (up to 35 total variables selected from preliminary linear models having Adjusted R-square higher than 10%) at each step has shown that lowest MSE has been achieved by a random forest with 9 variables at each step. See picture below. At this model, the correlation between predicted values and actual values from test data equals 95.05%.



Shortcomings

In linear models, in spite of dropping outliers, the final model still lacks normality and homoscedasticity assumptions which needs to be further addressed.

In data transformation for NAs, I have used average values or zero for numeric independent variables and “None” or most frequent values for categorical variables. I believe further data learning is necessary to better assign proper values to NAs.