Final Report about Prediction Sale Price of houses in Iowa

Indoo Park

March 24, 2019

1.Abstract:

The final report about prediction of sale price of houses in Ames, Iowa. The data set was given by Professor. Based on the given training data I built a model. To make the model I was needed to transform the missing vectors to character vector “None”. First model had many predictors which could violate the overfitting. For reducing predictors, I compared each predictors significances and got rid of low significant predictors. My final model got 92% of R-Squared with 22 predictors.

I predicted prices of house of testing data set with my final model and Kaggle calculated my testing R-Squared as 92%, my Kaggle rank is 3rd; however, that was the rank of my first submission with use of randomforest. I don’t know the exact rank for this testing model with use of the multiple linear regression way.

2. Introduction:

When a home buyer decide to buy a house or not, the most important factor is its price. If the buyer could get the data of houses in Iowa and have a best model for prediction the true value of a house, He or she could know the house is undervalued or overpriced. Then, he or she would decide to invest or not. Our goal is making the best model for prediction of sale price of houses in Ames, Iowa for a buyer to make the best decision. A buyer will buy after comparing with true value of the house and buy if the house is undervalued or not overpriced.

The train data has 81 explanatory variables for the 2500 different houses in Ames, Iowa. 81 variables are the factors to make their sale price such as “Lot size in square feet”, “Type of road access”, “Slope of property”, “Style of dwelling”, “year built”, “type of heating”, “garage size”, etc. Test data has 80 variables without sale price and 1500 different house. I will use explanatory variables where in train data with the multiple linear regression to make the best model for prediction of the sale price of houses in Ames, Iowa.

3. Methodology.

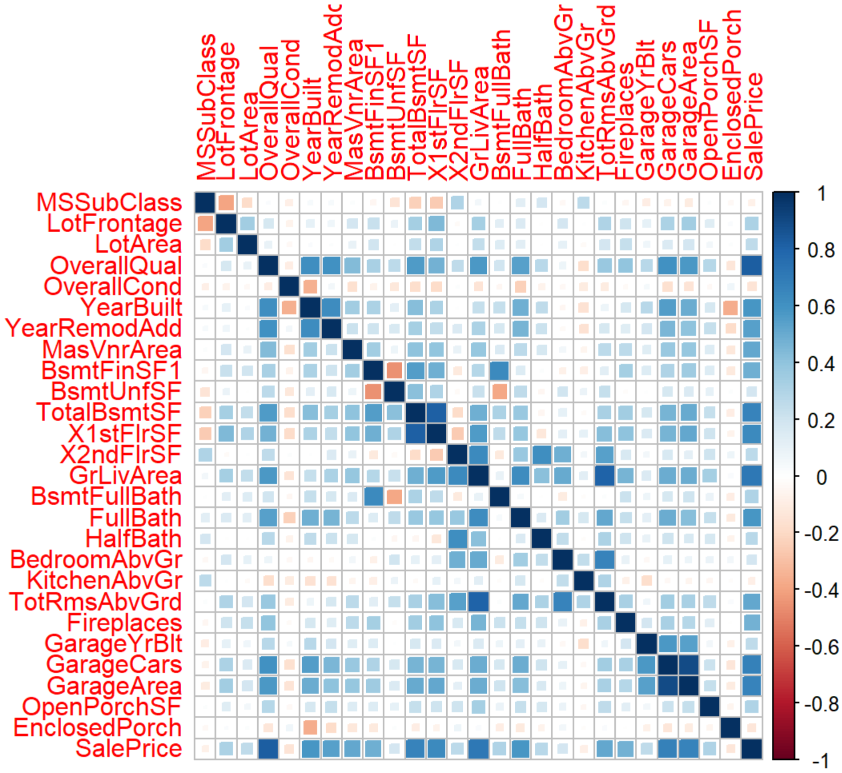
When I looked the data, the data contains qualitative data and quantitative data together. To make a multiple linear regression model, I needed to factorize the data. To facotorize the data I firstly find the missing values which were “NA” and changed them with mean or a character vector “None”.

```
correlations <- cor(na.omit(train1_num[,-1]))
```

```
#correlation between predictors and saleprice
correlations[,ncol(correlations)]
```

```
##    MSSubClass   LotFrontage       LotArea   OverallQual   OverallCond
##  -0.075853853   0.317892699   0.251100273   0.823249703  -0.142423570
##     YearBuilt  YearRemodAdd    MasVnrArea    BsmtFinSF1    BsmtFinSF2
##   0.582547075   0.545074904   0.513369484   0.488442485  -0.031713206
##     BsmtUnfSF   TotalBsmtSF     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##   0.183008458   0.663666238   0.634930276   0.269704623  -0.085478288
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##   0.719917951   0.308098784  -0.023264812   0.582260385   0.276858494
##  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd    Fireplaces   GarageYrBlt
##   0.186340834  -0.088461730   0.513398082   0.475297256   0.258888893
##    GarageCars    GarageArea    WoodDeckSF   OpenPorchSF EnclosedPorch
##   0.677912769   0.669700448   0.298294594   0.304447179  -0.131116815
##    X3SsnPorch   ScreenPorch      PoolArea       MiscVal        MoSold
##   0.003787636   0.133060359   0.018721808   0.032679322   0.021609086
##        YrSold     SalePrice
##  -0.013995073   1.000000000
```

```
corrplot(correlations, method="square")
```



I checked the correlations between predictors and sale price with corrlation function and their correlation plot. I selected high correlated predictors to make my first draft model m1.

```
m1 <- lm(SalePrice~MSSubClass+LotFrontage+LotArea+OverallQual+OverallCond+YearBuilt+YearRemodAdd+MasVnrArea+
BsmtFinSF1+BsmtFinSF2+BsmtUnfSF+TotalBsmtSF+X1stFlrSF+X2ndFlrSF+LowQualFinSF+GrLivArea+BsmtFullBath+BsmtHalf
Bath+FullBath+HalfBath+BedroomAbvGr+KitchenAbvGr+TotRmsAbvGrd+Fireplaces+GarageYrBlt+GarageCars+GarageArea,d
ata=train)
#summary(m1)
summary(lm(m1,train))$r.squared#r.squared
```

```
## [1] 0.8611272
```

It was the good start because I got 86 % of R-squared, which mean my model explained 86% of data. However, I still wanted to improve my r-squared to make the best model, which has lowest number of predictors with high r-squared.

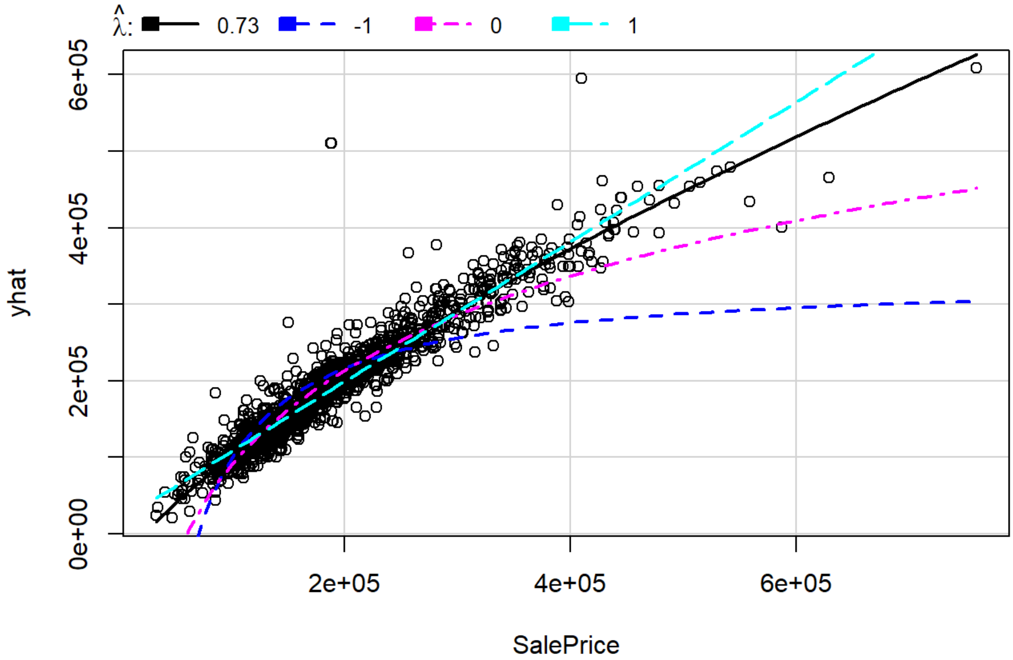
```
step(m1, direction = "backward", data=train)
```

```
#new model with predictors which have the lowest AIC
m2 <- lm(SalePrice~LotArea+OverallQual+OverallCond +YearRemodAdd + MasVnrArea + BsmtFinSF1+TotalBsmtSF+X1stF
lrSF+X2ndFlrSF+BsmtFullBath+Fireplaces+GarageArea+Neighborhood+Exterior1st+LandContour+LotConfig+GarageYrBlt
+BldgType+HouseStyle+RoofMatl+ExterQual+SaleCondition,data=train)
#summary(m2)
summary(lm(m2,train))$r.squared #r squared
```

```
## [1] 0.9141743
```

I used step function to reduce the number of predictors. The step function shows AIC, so I can easiliy choose the best predictors with the lowest AIC. Frome the result of step function which has the lowest AIC, I reduced some predictors and made second draft model m2. I checked my r-squared got improved to 91%.

```
#find lambda for trasformation
inverseResponsePlot(m2)
```

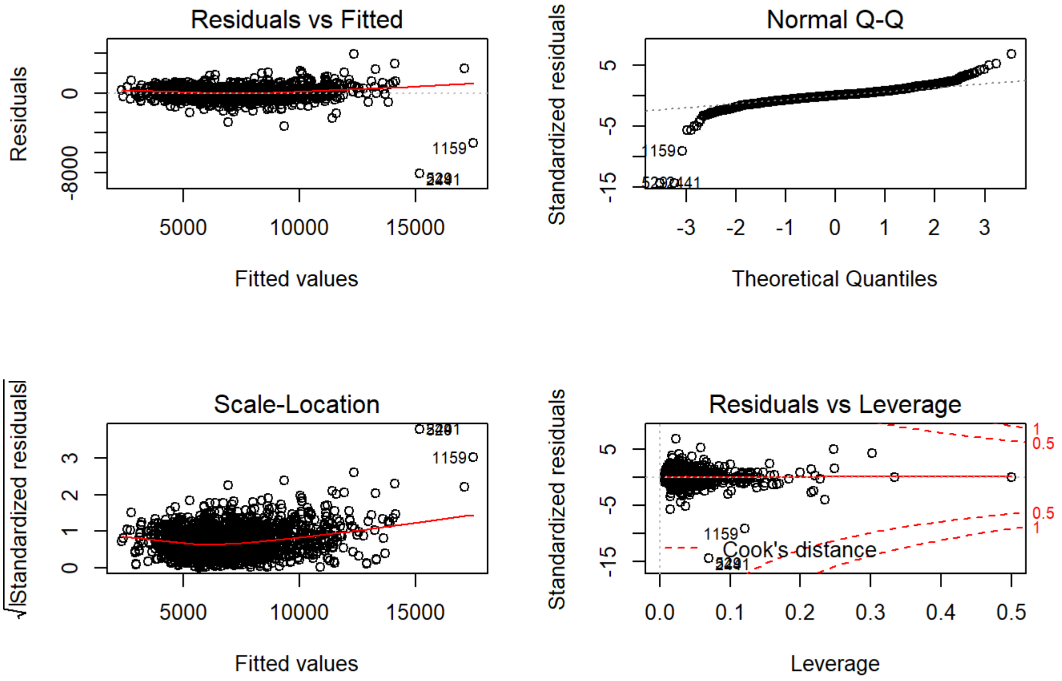


I needed to check the validity of my model. Since my RSE of my model is super high, I wanted to transform with the best lambda to check the plot. By use of inverseresponseplot function, I got the best lambda .73.

```
par(mfrow=c(2,2))
plot(m2)
```

```
## Warning: not plotting observations with leverage one:
##   701, 848, 1306

## Warning: not plotting observations with leverage one:
##   701, 848, 1306
```



First diagnostic graph shows almost flat line which represents the linearility of the model, and second graph shows that the points are placed along with the line which means the model has the normality. Third plot shows constant variance but not perfectly. Last graph shows that the model has leverage points. Fianlly, we can conclude that the model is almost valid base on the plots.

```
#check vif
v<-vif(m2)
vif.table <- data.frame((v[,3]))
vif.table
```

```
##                X.v...3..
## LotArea        1.251140
## OverallQual    2.101429
## OverallCond    1.266244
## YearRemodAdd   1.671926
## MasVnrArea     1.353914
## BsmtFinSF1     1.630466
## TotalBsmtSF    2.121818
## X1stFlrSF      2.319026
## X2ndFlrSF      2.767636
## BsmtFullBath   1.404107
## Fireplaces     1.321299
## GarageArea     1.768816
## Neighborhood   1.135574
## Exterior1st    1.109606
## LandContour    1.147567
## LotConfig      1.043953
## GarageYrBlt    1.387264
## BldgType       1.244240
## HouseStyle     1.256765
## RoofMatl       1.064921
## ExterQual      1.347054
## SaleCondition  1.093317
```

To make the model more valid, I checked Variance inflation factor (VIF). I got rid of some predictors which have more than 5 VIF number.

```
anova<- anova(m2)
anova
```

```
od<-order(anova[,"Pr(>F)"], decreasing = TRUE)
anova[od,]
```

```
## Analysis of Variance Table
##
## Response: (SalePrice)^0.73
##                 Df     Sum Sq    Mean Sq    F value    Pr(>F)    
## GarageYrBlt      1    5565349    5565349    16.3705 5.373e-05 ***
## LotConfig        4    9437913    2359478     6.9404 1.493e-05 ***
## SaleCondition    5   11098776    2219755     6.5294 4.842e-06 ***
## Fireplaces       1   11186741   11186741    32.9058 1.088e-08 ***
## HouseStyle       7   21978786    3139827     9.2358 2.541e-11 ***
## Exterior1st     14   28640762    2045769     6.0176 7.776e-12 ***
## BsmtFullBath     1   24268941   24268941    71.3872 < 2.2e-16 ***
## LandContour      3   29245616    9748539    28.6754 < 2.2e-16 ***
## OverallCond      1   34291969   34291969   100.8700 < 2.2e-16 ***
## ExterQual        3   40736431   13578810    39.9421 < 2.2e-16 ***
## BldgType         4   53464637   13366159    39.3166 < 2.2e-16 ***
## YearRemodAdd     1   90224133   90224133   265.3947 < 2.2e-16 ***
## GarageArea       1  116714211  116714211   343.3154 < 2.2e-16 ***
## TotalBsmtSF      1  128516688  128516688   378.0324 < 2.2e-16 ***
## X1stFlrSF        1  130650676  130650676   384.3095 < 2.2e-16 ***
## RoofMatl         5  161882264   32376453    95.2355 < 2.2e-16 ***
## Neighborhood    24  244811941   10200498    30.0048 < 2.2e-16 ***
## MasVnrArea       1  210278721  210278721   618.5358 < 2.2e-16 ***
## BsmtFinSF1       1  298973056  298973056   879.4305 < 2.2e-16 ***
## X2ndFlrSF        1  415713475  415713475  1222.8230 < 2.2e-16 ***
## LotArea          1  662774376  662774376  1949.5537 < 2.2e-16 ***
## OverallQual      1 7113044479 7113044479 20923.0508 < 2.2e-16 ***
## Residuals     2414  820668529     339962                         
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#summary(m2)
summary(lm(m2,train))$r.squared
```

```
## [1] 0.9230443
```

Also, I checked the p-value of each predictors and I got rid of some which have p-value which is more than 0.05. Finally, I made my final model with 92% of R-squared.

```
prediction <- predict(m2, test)
summary(prediction)

prediction <- prediction^(1/0.73)
prediction <- data.frame(Ob = 1:1500, SalePrice = prediction)
```

By using of predict function I could predict the saleprice of test data set. Since used transformed model, I retransformed the result to make the price normal.

4. Results.

I was able to reduce the number of predictors from 82 to 22, and got a 92% of r-squared. I made a submission on Kaggle, and it calculated my testing r-squared of 92%. Since the professor’s threshold for r-squared is 83%, I conclude that I made a very good result. In the other word my model has 92% accuracy.

5. Discussion.

Not all of the 82 variables of Train data set is important to predict the sale price because there are a few variables have correlation with sale price. In other words, there are significant variables and insignificant variables. My final model has 22 variables, that correlate with sale price, and I got 92% r-squared or 92% accuracy. Also the testing r-squared of 92% means my final model explains 92% of variability of the response data around its mean. By checking VIF, my variables are not violating multicorrinearlity. The checking if there is a multicorrinearlity between selected variables is the most important requirement for the model.

6.Limitations and Conclusion

I used only 22 variables to predict new pirce. This reduction makes program runs faster and efficiently better for the buyer. Although my final model got 92% of high r-squared number, the model is not the perfect model because it still has 8% of error. To make the better model I will be needed to delete all of the bad leverage or find other best combination of predictors or use other classification techniques like randomforest, knn, gradient boosting. Next quarter, I will may be able to make this model much better.

I recommend the buyer use my final model to predict the sale price because 92% accuracy is high enought to use. If a buyer uses my final model for his investment, he would be able to recognize which houses are undervalued or which houses are overpriced.

7.Reference.

Nau Robert. “Statistical Forecasting: notes on regression and time series analysis”. https://people.duke.edu/~rnau/411home.htm.

Kaggle. “Predicting Sale Price of Houses in Iowa”. https://www.kaggle.com/c/stat101ahouseprice.

Epidemiol J Nepal. “Understanding Significance and P-Values”, 2016 Mar. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4850233/