# Assignment 5: Automated Variable selection, multicollinearity and predictive modeling.

*Sri Seshadri*

*7/19/2017*

## 1. Introduction

## 2. Sample definition

It is assumed that typical home buyers are those that move from apartments to single family or town homes. Also apartments are less likely to be sold to individuals as they remain holdings of owners for rental income. Single family and town homes belong to "Residential Low density" (RL) zoning classification in the city of Ames. Data belonging to only to the RL zone is considered for analysis and model development. Also, it is assumed that typical homes have paved streets for access and above grade living area greater than 800 square feet. Sales data belonging to homes that were sold in abnormal conditions such as trade in, foreclosure or short sale are not included in the analysis. Also, sales between family members, sale of adjoining lot, linked properties are omitted from the data. Homes with no basements are excluded from the analysis at this time. Homes with living area greater than 4000 square feet and garage area greater than 1000 square feet were identified as outliers and are therefore removed from the analysis. Table 1 shows the waterfall of the data not included in the data and the eligible samples.

Table 1: Drop waterfall

| DropCondition | counts |
|---|---:|
| 01: Not LowDensityZone | 657 |
| 02: Not Normal/Partial Sale | 189 |
| 03: Street Not Paved | 3 |
| 04: Less than 800 SqFt | 41 |
| 05: No Basement | 48 |
| 06: Greater 4000 sqft living Area - Influence Points | 4 |
| 07:Garage area greater than 1000 sqft - Influence points | 23 |
| 99: Eligible Sample | 1965 |

### 2.1 Predictor variables of interest for modelling

The following variables in the data were deemed to be of interest for model building. The choice of parameters was based upon intial Exploratory Data Analysis (EDA) and subject matter expertise. The categorical variables are coded into indicator variables for model building purposes, the tables 3, 4, 5 and 6 show the description of the indicator variables.

Table 2: Predictors of interest

| | | |
|---|---|---|
| LotArea | BsmtFullBath | MoSold |
| LotConfig | BsmtHalfBath | YrSold |
| Neighborhood | FullBath | SaleCondition |
| BldgType | HalfBath | FirstFlrSF |
| OverallCond | BedroomAbvGr | SecondFlrSF |
| YearRemodel | KitchenQual | OverallQual |

| | |
|---|---|
| TotalBsmtSF | TotRmsAbvGrd |
| GrLivArea | GarageArea |

The following tables describe the indicator variables:

Table 3: Neighborhood tiers,base category > 90

| Tier | Price.per.sq.ft |
|---|---|
| 1 | <= 60 |
| 2 | > 60 and <= 70 |
| 3 | > 70 and <= 80 |
| 4 | > 80 and <= 90 |

Table 4: Lot configuration indicator variables; base category: Inside Lot

| Indicator | Description |
|---|---|
| CornerLot | Corner lot |
| CulDSac | CulDSac Lot |
| Frontal2 | 2 frontal lot |
| Frontal3 | 3 frontal lot |

Table 5: Building type indicator variables; base category: single family

| Indicator | Decription |
|---|---|
| TwnhsE | Townhouse |
| Twnhs | Twin house |
| Duplex | Duplex |
| fam2 | 2 family conversion |

Table 6: Kitchen Quality indicator variables; base category: poor

| Indicator | Decription |
|---|---|
| KTA | Typical/Average |
| KGD | Good |
| KEx | Excellent |
| KFa | Fair |

### 2.2 Training and validation samples.

From the eligible samples, 70% of the data is randomly sampled to be used as the dataset for model development. This dataset would be refered to as training dataset. The remaining 30% is used as the validation set to evaluate the model performance of predicting sale price on data that is outside the training set. Table 7 shows the split of the total eligible samples.

Table 7: Training and Validation sampling

| Data | Samples |
|---|---:|
| Training set | 1382 |
| Validation set | 583 |
| Total | 1965 |

## 3. Model identification by Automated Variable Selection

Attribute variables such as “TotalSQFT”; sum of basement square feet and above grade living area and “QualityIndex”; product of overall quality and overall condition has been created for use as predictor variables in model building. With the attribute variables and the indicator variables created, below is a list of predictor variables that will be used for automated variable selection and modelling.

Table 8: Predictors for linear regression models

| | | | |
|---|---|---|---|
| LotArea | QualityIndex | Tier3 | TwnhsE |
| YearRemodel | TotRmsAbvGrd | Tier4 | Twnhs |
| TotalBsmtSF | GarageArea | PartialSaleCond | Duplex |
| GrLivArea | YearMonthSold | CornerLot | KTA |
| TotalBath | Tier1 | CulDSac | KGD |
| TotalSQFT | Tier2 | Frontal2 | KEx |

### 3.1. Full model

A full model with all the predictors used is fit as an exhaustive search. This will be used as an upper boundry for automated variable selection processes; forward selection and as a starting condition in the backward selection process. The full model and its fit is shown below:

**SalePrice = 1350637.8 + 0.39 * LotArea + 111.98 * YearRemodel - 29.47 * TotalBsmtSF - 16.79 * GrLivArea + 8713.08 * TotalBath + 73.3 * TotalSQFT + 1420.94 * QualityIndex - 2034.3 * TotRmsAbvGrd + 42.02 * GarageArea - 775.31 * YearMonthSold - 59226.84 * Tier1 - 54491.76 * Tier2 - 42884.83 * Tier3 - 18664.53 * Tier4 + 12306.86 * PartialSaleCond - 2414.41 * CornerLot + 5911.92 * CulDSac - 3085.71 * Frontal2 - 16913.19 * TwnhsE - 19016.5 * Twnhs - 39824.09 * Duplex - 2600.99 * KTA + 393.06 * KGD + 48661.14 * KEx**

```
##
## Call:
## lm(formula = SalePrice ~ ., data = train.Clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -140178  -12906    -355   12866  186447
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.351e+06  1.065e+06   1.269  0.20473
## LotArea          3.896e-01  8.224e-02   4.737 2.39e-06 ***
## YearRemodel      1.120e+02  5.474e+01   2.046  0.04098 *
## TotalBsmtSF     -2.947e+01  1.729e+01  -1.704  0.08857 .
## GrLivArea       -1.679e+01  1.699e+01  -0.988  0.32322
## TotalBath        8.713e+03  1.058e+03   8.239 4.05e-16 ***
## TotalSQFT        7.330e+01  1.708e+01   4.291 1.91e-05 ***
## QualityIndex     1.421e+03  1.011e+02  14.051  < 2e-16 ***
## TotRmsAbvGrd    -2.034e+03  8.329e+02  -2.442  0.01472 *
## GarageArea       4.202e+01  4.840e+00   8.681  < 2e-16 ***
## YearMonthSold   -7.753e+02  5.318e+02  -1.458  0.14512
## Tier1           -5.923e+04  5.536e+03 -10.699  < 2e-16 ***
## Tier2           -5.449e+04  4.944e+03 -11.023  < 2e-16 ***
## Tier3           -4.288e+04  4.723e+03  -9.080  < 2e-16 ***
## Tier4           -1.866e+04  4.668e+03  -3.999 6.71e-05 ***
## PartialSaleCond  1.231e+04  2.657e+03   4.633 3.96e-06 ***
## CornerLot       -2.414e+03  1.834e+03  -1.316  0.18832
## CulDSac          5.912e+03  2.535e+03   2.332  0.01982 *
## Frontal2        -3.086e+03  3.873e+03  -0.797  0.42572
## TwnhsE          -1.691e+04  3.324e+03  -5.089 4.11e-07 ***
## Twnhs           -1.902e+04  6.040e+03  -3.148  0.00168 **
## Duplex          -3.982e+04  4.392e+03  -9.067  < 2e-16 ***
## KTA             -2.601e+03  6.245e+03  -0.416  0.67712
## KGD              3.931e+02  6.508e+03   0.060  0.95185
## KEx              4.866e+04  7.261e+03   6.702 3.01e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24940 on 1357 degrees of freedom
## Multiple R-squared:  0.8932, Adjusted R-squared:  0.8913
## F-statistic:   473 on 24 and 1357 DF,  p-value: < 2.2e-16
```

### 3.2. Intercept only model as start for forward variable selection

An intercept only model is used as a lower boundary (no predictors used) condition for forward selection.

$$\text{Sale Price} = 195796.63$$

```
##
## Call:
## lm(formula = SalePrice ~ 1, data = train.Clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -137297  -52797  -19797   32203  419203
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   195797       2035   96.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75640 on 1381 degrees of freedom
```

### 3.3. Simple linear regression as starting condition for stepwise variable selection

A simple linear regression model with total square feet as predictor is used as a starting point for stepwise variable selection.Below is the model and fit:

*SalePrice = -41869.7 + 88.28 * TotalSQFT*

```
##
## Call:
## lm(formula = SalePrice ~ TotalSQFT, data = train.Clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197369  -20019    1018   20624  215020
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -41869.698   4369.978  -9.581   <2e-16 ***
## TotalSQFT       88.282      1.569  56.270   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41690 on 1380 degrees of freedom
## Multiple R-squared:  0.6965, Adjusted R-squared:  0.6962
## F-statistic:  3166 on 1 and 1380 DF,  p-value: < 2.2e-16
```

## 3.4. Forward variable selection.

A forward selection method is used for selecting variables for linear regression fit. The following model was fit by the automated forward selection by AIC.

**SalePrice = 1434891.27 + 56.69 * TotalSQFT + 1412.82 * QualityIndex + 48321.03 * KEx + 41.71 * GarageArea - 54628.67 * Tier2 - 39765.64 * Duplex + 8784.07 * TotalBath - 18693.05 * Tier4 - 59496.11 * Tier1 - 43011.63 * Tier3 - 16782.73 * TwnhsE + 12253.21 * PartialSaleCond + 0.38 * LotArea - 19416.82 * Twnhs + 6525.48 * CulDSac + 116.3 * YearRemodel - 12.7 * TotalBsmtSF - 2162.38 * TotRmsAbvGrd - 2955.09 * KTA - 821.27 * YearMonthSold**

```
##
## Call:
## lm(formula = SalePrice ~ TotalSQFT + QualityIndex + KEx + GarageArea +
##     Tier2 + Duplex + TotalBath + Tier4 + Tier1 + Tier3 + TwnhsE +
##     PartialSaleCond + LotArea + Twnhs + CulDSac + YearRemodel +
##     TotalBsmtSF + TotRmsAbvGrd + KTA + YearMonthSold, data = train.Clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -139711  -13263    -437   12742  186642
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.435e+06  1.062e+06   1.351  0.17704
## TotalSQFT        5.669e+01  3.105e+00  18.260  < 2e-16 ***
## QualityIndex     1.413e+03  1.000e+02  14.123  < 2e-16 ***
## KEx              4.832e+04  3.008e+03  16.062  < 2e-16 ***
## GarageArea       4.171e+01  4.800e+00   8.689  < 2e-16 ***
## Tier2           -5.463e+04  4.941e+03 -11.056  < 2e-16 ***
## Duplex          -3.977e+04  4.381e+03  -9.078  < 2e-16 ***
## TotalBath        8.784e+03  1.056e+03   8.317  < 2e-16 ***
## Tier4           -1.869e+04  4.666e+03  -4.006 6.50e-05 ***
## Tier1           -5.950e+04  5.522e+03 -10.774  < 2e-16 ***
## Tier3           -4.301e+04  4.719e+03  -9.114  < 2e-16 ***
## TwnhsE          -1.678e+04  3.314e+03  -5.064 4.68e-07 ***
## PartialSaleCond  1.225e+04  2.655e+03   4.615 4.29e-06 ***
## LotArea          3.839e-01  8.209e-02   4.676 3.21e-06 ***
## Twnhs           -1.942e+04  5.959e+03  -3.258  0.00115 **
## CulDSac          6.525e+03  2.501e+03   2.609  0.00918 **
## YearRemodel      1.163e+02  5.400e+01   2.154  0.03143 *
## TotalBsmtSF     -1.270e+01  4.186e+00  -3.033  0.00247 **
## TotRmsAbvGrd    -2.162e+03  8.221e+02  -2.630  0.00862 **
## KTA             -2.955e+03  1.907e+03  -1.549  0.12153
## YearMonthSold   -8.213e+02  5.306e+02  -1.548  0.12193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24930 on 1361 degrees of freedom
## Multiple R-squared:  0.893,  Adjusted R-squared:  0.8914
## F-statistic: 567.8 on 20 and 1361 DF,  p-value: < 2.2e-16
```

### 3.5 Backward elimination selection.

The following model is a result of a backward elimination selection, with full model as its starting model.

**SalePrice = 1434891.27 + 0.38 * LotArea + 116.3 * YearRemodel - 12.7 * TotalBsmtSF + 8784.07 * TotalBath + 56.69 * TotalSQFT + 1412.82 * QualityIndex - 2162.38 * TotRmsAbvGrd + 41.71 * GarageArea - 821.27 * YearMonthSold - 59496.11 * Tier1 - 54628.67 * Tier2 - 43011.63 * Tier3 - 18693.05 * Tier4 + 12253.21 * PartialSaleCond + 6525.48 * CulDSac - 16782.73 * TwnhsE - 19416.82 * Twnhs - 39765.64 * Duplex - 2955.09 * KTA + 48321.03 * KEx**

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + YearRemodel + TotalBsmtSF +
##     TotalBath + TotalSQFT + QualityIndex + TotRmsAbvGrd + GarageArea +
##     YearMonthSold + Tier1 + Tier2 + Tier3 + Tier4 + PartialSaleCond +
##     CulDSac + TwnhsE + Twnhs + Duplex + KTA + KEx, data = train.Clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -139711  -13263    -437   12742  186642
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.435e+06  1.062e+06   1.351  0.17704
## LotArea          3.839e-01  8.209e-02   4.676 3.21e-06 ***
## YearRemodel      1.163e+02  5.400e+01   2.154  0.03143 *
## TotalBsmtSF     -1.270e+01  4.186e+00  -3.033  0.00247 **
## TotalBath        8.784e+03  1.056e+03   8.317  < 2e-16 ***
## TotalSQFT        5.669e+01  3.105e+00  18.260  < 2e-16 ***
## QualityIndex     1.413e+03  1.000e+02  14.123  < 2e-16 ***
## TotRmsAbvGrd    -2.162e+03  8.221e+02  -2.630  0.00862 **
## GarageArea       4.171e+01  4.800e+00   8.689  < 2e-16 ***
## YearMonthSold   -8.213e+02  5.306e+02  -1.548  0.12193
## Tier1           -5.950e+04  5.522e+03 -10.774  < 2e-16 ***
## Tier2           -5.463e+04  4.941e+03 -11.056  < 2e-16 ***
## Tier3           -4.301e+04  4.719e+03  -9.114  < 2e-16 ***
## Tier4           -1.869e+04  4.666e+03  -4.006 6.50e-05 ***
## PartialSaleCond  1.225e+04  2.655e+03   4.615 4.29e-06 ***
## CulDSac          6.525e+03  2.501e+03   2.609  0.00918 **
## TwnhsE          -1.678e+04  3.314e+03  -5.064 4.68e-07 ***
## Twnhs           -1.942e+04  5.959e+03  -3.258  0.00115 **
## Duplex          -3.977e+04  4.381e+03  -9.078  < 2e-16 ***
## KTA             -2.955e+03  1.907e+03  -1.549  0.12153
## KEx              4.832e+04  3.008e+03  16.062  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24930 on 1361 degrees of freedom
## Multiple R-squared:  0.893,  Adjusted R-squared:  0.8914
## F-statistic: 567.8 on 20 and 1361 DF,  p-value: < 2.2e-16
```

## 3.6. Stepwise regression method.

The following model is a result of stepwise variable selection with full model and simple linear regression model with Total square feet as a predictor as boundary conditions.

**SalePrice = 1434891.27 + 56.69 * TotalSQFT + 1412.82 * QualityIndex + 48321.03 * KEx + 41.71 * GarageArea - 54628.67 * Tier2 - 39765.64 * Duplex + 8784.07 * TotalBath - 18693.05 * Tier4 - 59496.11 * Tier1 - 43011.63 * Tier3 - 16782.73 * TwnhsE + 12253.21 * PartialSaleCond + 0.38 * LotArea - 19416.82 * Twnhs + 6525.48 * CulDSac + 116.3 * YearRemodel - 12.7 * TotalBsmtSF - 2162.38 * TotRmsAbvGrd - 2955.09 * KTA - 821.27 * YearMonthSold**

```
##
## Call:
## lm(formula = SalePrice ~ TotalSQFT + QualityIndex + KEx + GarageArea +
##     Tier2 + Duplex + TotalBath + Tier4 + Tier1 + Tier3 + TwnhsE +
##     PartialSaleCond + LotArea + Twnhs + CulDSac + YearRemodel +
##     TotalBsmtSF + TotRmsAbvGrd + KTA + YearMonthSold, data = train.Clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -139711  -13263    -437   12742  186642
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.435e+06  1.062e+06   1.351  0.17704
## TotalSQFT        5.669e+01  3.105e+00  18.260  < 2e-16 ***
## QualityIndex     1.413e+03  1.000e+02  14.123  < 2e-16 ***
## KEx              4.832e+04  3.008e+03  16.062  < 2e-16 ***
## GarageArea       4.171e+01  4.800e+00   8.689  < 2e-16 ***
## Tier2           -5.463e+04  4.941e+03 -11.056  < 2e-16 ***
## Duplex          -3.977e+04  4.381e+03  -9.078  < 2e-16 ***
## TotalBath        8.784e+03  1.056e+03   8.317  < 2e-16 ***
## Tier4           -1.869e+04  4.666e+03  -4.006 6.50e-05 ***
## Tier1           -5.950e+04  5.522e+03 -10.774  < 2e-16 ***
## Tier3           -4.301e+04  4.719e+03  -9.114  < 2e-16 ***
## TwnhsE          -1.678e+04  3.314e+03  -5.064 4.68e-07 ***
## PartialSaleCond  1.225e+04  2.655e+03   4.615 4.29e-06 ***
## LotArea          3.839e-01  8.209e-02   4.676 3.21e-06 ***
## Twnhs           -1.942e+04  5.959e+03  -3.258  0.00115 **
## CulDSac          6.525e+03  2.501e+03   2.609  0.00918 **
## YearRemodel      1.163e+02  5.400e+01   2.154  0.03143 *
## TotalBsmtSF     -1.270e+01  4.186e+00  -3.033  0.00247 **
## TotRmsAbvGrd    -2.162e+03  8.221e+02  -2.630  0.00862 **
## KTA             -2.955e+03  1.907e+03  -1.549  0.12153
## YearMonthSold   -8.213e+02  5.306e+02  -1.548  0.12193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24930 on 1361 degrees of freedom
## Multiple R-squared:  0.893,  Adjusted R-squared:  0.8914
## F-statistic: 567.8 on 20 and 1361 DF,  p-value: < 2.2e-16
```

### 3.7. Junk Model

The attempt here is to verify the intuition of how correlated predictor variables perform. And how they compare to models from automated variable selection. The predictors chosen are OverallQual,OverallCond, QualityIndex, GrLivArea and TotalSQFT. QualityIndex is a linear combination of OverallQual and OverllCond. TotalSQFT is a result of a linear combination of variables that include GrLivArea. Below is the junk model and its fit.

**SalePrice = -264550.51 + 44886.27 * OverallQual + 25559.84 * OverallCond - 3492.08 * QualityIndex + 13.7 * GrLivArea + 50.41 * TotalSQFT**

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + OverallCond + QualityIndex +
##     GrLivArea + TotalSQFT, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -114864  -19020    -142   17350  191358
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.646e+05  2.220e+04 -11.916  < 2e-16 ***
## OverallQual   4.489e+04  3.660e+03  12.265  < 2e-16 ***
## OverallCond   2.556e+04  4.109e+03   6.220 6.58e-10 ***
## QualityIndex -3.492e+03  6.899e+02  -5.062 4.71e-07 ***
## GrLivArea     1.370e+01  3.525e+00   3.886 0.000107 ***
## TotalSQFT     5.041e+01  2.580e+00  19.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31340 on 1376 degrees of freedom
## Multiple R-squared:  0.8289, Adjusted R-squared:  0.8283
## F-statistic:  1333 on 5 and 1376 DF,  p-value: < 2.2e-16
```

Table 9: VIF - Top 5 my model

| Model | Predictors | VIF |
|---|---|---|
| Junk | QualityIndex | 45.61 |
| | OverallQual | 32.30 |
| | OverallCond | 24.31 |
| | TotalSQFT | 4.78 |
| | GrLivArea | 4.06 |
| Forward Selection | Tier2 | 12.86 |
| | Tier3 | 11.23 |
| | TotalSQFT | 10.95 |
| | Tier4 | 6.81 |
| | TotalBsmtSF | 5.64 |
| Backward Elimination | Tier2 | 12.86 |
| | Tier3 | 11.23 |
| | TotalSQFT | 10.95 |
| | Tier4 | 6.81 |
| | TotalBsmtSF | 5.64 |
| Stepwise | Tier2 | 12.86 |
| | Tier3 | 11.23 |
| | TotalSQFT | 10.95 |
| | Tier4 | 6.81 |
| | TotalBsmtSF | 5.64 |

### 3.7.1 Variance Inflation Factors (VIF)

Having known that the predictors in the junk model are correlated, it would be interesting to compare correlation amongst predictors in the models from automated variable selection methods against that of the junk model. Table 9 shows the top 5 highest VIFs by models.

It is not surprising that the highest VIFs are from the junk model. However we see neighborhood Tiers 2 and 3 have high VIF ( $> 10$). This is not surprising either as they are indicator(dummy) variables. The indicator variables are correlated amongst themselves and can be modelled as linear combinations of each other. Thereby yielding a high coefficient of determintion when indicator variables are regressed by other predictors, causing an inflated VIF.

## 4. Model Comparison

Table 10: Model comparison

| model | adj.r.squared | AIC | BIC | MSE | MAE |
|---|---|---|---|---|---|
| Junk | 0.8283004 | 32544.92 | 32581.54 | 978115255 | 23151.42 |
| Forward Selection | 0.8913980 | 31926.74 | 32041.82 | 611924810 | 17596.46 |
| Backward Elimination | 0.8913980 | 31926.74 | 32041.82 | 611924810 | 17596.46 |
| Stepwise | 0.8913980 | 31926.74 | 32041.82 | 611924810 | 17596.46 |