

# 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 initial 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	Description
TwnhsE	Townhouse
Twnhs	Twin house
Duplex	Duplex
fam2	2 family conversion

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

Indicator	Description
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 referred 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	1367
Validation set	598
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:

$\text{SalePrice} = 321607.84 + 0.53 * \text{LotArea} + 142.21 * \text{YearRemodel} - 23.41 * \text{TotalBsmtSF} - 9.86 * \text{GrLivArea} + 8818.06 * \text{TotalBath} + 67.42 * \text{TotalSQFT} + 1462.05 * \text{QualityIndex} - 1692.59 * \text{TotRmsAbvGrd} + 41.48 * \text{GarageArea} - 293.12 * \text{YearMonthSold} - 64167.8 * \text{Tier1} - 58735.13 * \text{Tier2} - 46878.43 * \text{Tier3} - 24076.41 * \text{Tier4} + 12878.55 * \text{PartialSaleCond} - 3187.7 * \text{CornerLot} + 7416.17 * \text{CulDSac} - 3193.34 * \text{Frontal2} - 12735.7 * \text{TwnhsE} - 20269.07 * \text{Twnhs} - 42869.14 * \text{Duplex} - 1911.98 * \text{KTA} - 1893.94 * \text{KGD} + 41586.62 * \text{KEx}$

```
##
## Call:
## lm(formula = SalePrice ~ ., data = train.Clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -147695  -12828    -721   12040  201006
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.216e+05  1.015e+06   0.317  0.751387
## LotArea      5.340e-01  9.507e-02   5.617  2.36e-08 ***
## YearRemodel  1.422e+02  5.219e+01   2.725  0.006520 **
## TotalBsmtSF -2.341e+01  1.557e+01  -1.503  0.133029
## GrLivArea   -9.858e+00  1.528e+01  -0.645  0.518953
## TotalBath    8.818e+03  1.049e+03   8.409  < 2e-16 ***
```

```
## TotalsQFT      6.742e+01  1.535e+01  4.393 1.21e-05 ***
## QualityIndex   1.462e+03  1.020e+02 14.327 < 2e-16 ***
## TotRmsAbvGrd  -1.693e+03  8.214e+02 -2.061 0.039525 *
## GarageArea     4.148e+01  4.664e+00  8.894 < 2e-16 ***
## YearMonthSold -2.931e+02  5.062e+02 -0.579 0.562622
## Tier1          -6.417e+04  5.671e+03 -11.315 < 2e-16 ***
## Tier2          -5.874e+04  5.116e+03 -11.481 < 2e-16 ***
## Tier3          -4.688e+04  4.911e+03 -9.546 < 2e-16 ***
## Tier4          -2.408e+04  4.858e+03 -4.956 8.12e-07 ***
## PartialSaleCond 1.288e+04  2.565e+03  5.021 5.83e-07 ***
## CornerLot      -3.188e+03  1.812e+03 -1.760 0.078693 .
## CulDSac        7.416e+03  2.548e+03  2.911 0.003661 **
## Frontal2       -3.193e+03  3.520e+03 -0.907 0.364408
## TwnhsE         -1.274e+04  3.214e+03 -3.962 7.81e-05 ***
## Twnhs          -2.027e+04  5.475e+03 -3.702 0.000222 ***
## Duplex         -4.287e+04  4.515e+03 -9.494 < 2e-16 ***
## KTA            -1.912e+03  5.616e+03 -0.340 0.733564
## KGD            -1.894e+03  5.868e+03 -0.323 0.746940
## KEx            4.159e+04  6.626e+03  6.276 4.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24210 on 1342 degrees of freedom
## Multiple R-squared:  0.8982, Adjusted R-squared:  0.8963
## F-statistic: 493.1 on 24 and 1342 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.

$$\text{Sale Price} = 195734.7$$

```
##
## Call:
## lm(formula = SalePrice ~ 1, data = train.Clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -137235  -52510  -19735   32765  429265
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   195735      2034    96.25 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75190 on 1366 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:

$$\text{SalePrice} = -41801.44 + 88.57 * \text{TotalSQFT}$$

```
##
## Call:
## lm(formula = SalePrice ~ TotalSQFT, data = train.Clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -198690  -20278    1405   19810  237688
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -41801.440   4376.085  -9.552  <2e-16 ***
## TotalSQFT      88.575      1.578   56.145  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41350 on 1365 degrees of freedom
## Multiple R-squared:  0.6978, Adjusted R-squared:  0.6976
## F-statistic: 3152 on 1 and 1365 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.

$$\begin{aligned} \text{SalePrice} = & -264173.38 + 57.95 * \text{TotalSQFT} + 1461.02 * \text{QualityIndex} + 43318.45 * \text{KEx} + \\ & 41.56 * \text{GarageArea} - 58888.49 * \text{Tier2} - 42874.6 * \text{Duplex} + 8797.95 * \text{TotalBath} - 64301.63 * \\ & \text{Tier1} - 47087.15 * \text{Tier3} + 0.53 * \text{LotArea} + 13237.91 * \text{PartialSaleCond} - 24196.97 * \text{Tier4} - \\ & 12788.64 * \text{TwnhsE} - 20965.71 * \text{Twnhs} + 7597.53 * \text{CulDSac} + 140.02 * \text{YearRemodel} - 13.93 \\ & * \text{TotalBsmtSF} - 1818.83 * \text{TotRmsAbvGrd} - 3002.58 * \text{CornerLot} \end{aligned}$$

```
##
## Call:
## lm(formula = SalePrice ~ TotalSQFT + QualityIndex + KEx + GarageArea +
##      Tier2 + Duplex + TotalBath + Tier1 + Tier3 + LotArea + PartialSaleCond +
##      Tier4 + TwnhsE + Twnhs + CulDSac + YearRemodel + TotalBsmtSF +
##      TotRmsAbvGrd + CornerLot, data = train.Clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -147708  -12715    -680   12197  200622
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.642e+05  9.522e+04  -2.774  0.005608 **
## TotalSQFT      5.795e+01  3.069e+00  18.884  < 2e-16 ***
## QualityIndex   1.461e+03  1.005e+02  14.533  < 2e-16 ***
## KEx           4.332e+04  2.900e+03  14.939  < 2e-16 ***
## GarageArea     4.156e+01  4.622e+00   8.991  < 2e-16 ***
```

```
## Tier2          -5.889e+04  5.098e+03 -11.551 < 2e-16 ***
## Duplex         -4.287e+04  4.462e+03  -9.609 < 2e-16 ***
## TotalBath       8.798e+03  1.046e+03   8.408 < 2e-16 ***
## Tier1          -6.430e+04  5.643e+03 -11.395 < 2e-16 ***
## Tier3          -4.709e+04  4.902e+03  -9.607 < 2e-16 ***
## LotArea         5.298e-01  9.467e-02   5.596 2.66e-08 ***
## PartialSaleCond 1.324e+04  2.497e+03   5.301 1.34e-07 ***
## Tier4          -2.420e+04  4.850e+03  -4.989 6.86e-07 ***
## TwnhsE         -1.279e+04  3.198e+03  -3.999 6.70e-05 ***
## Twnhs          -2.097e+04  5.437e+03  -3.856 0.000121 ***
## CulDSac         7.598e+03  2.532e+03   3.001 0.002744 **
## YearRemodel     1.400e+02  4.845e+01   2.890 0.003912 **
## TotalBsmtSF    -1.393e+01  4.076e+00  -3.418 0.000649 ***
## TotRmsAbvGrd   -1.819e+03  8.052e+02  -2.259 0.024048 *
## CornerLot      -3.003e+03  1.800e+03  -1.669 0.095440 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24180 on 1347 degrees of freedom
## Multiple R-squared:  0.898, Adjusted R-squared:  0.8966
## F-statistic: 624.3 on 19 and 1347 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 = -264173.38 + 0.53 \* LotArea + 140.02 \* YearRemodel - 13.93 \* TotalBsmtSF + 8797.95 \* TotalBath + 57.95 \* TotalSQFT + 1461.02 \* QualityIndex - 1818.83 \* TotRmsAbvGrd + 41.56 \* GarageArea - 64301.63 \* Tier1 - 58888.49 \* Tier2 - 47087.15 \* Tier3 - 24196.97 \* Tier4 + 13237.91 \* PartialSaleCond - 3002.58 \* CornerLot + 7597.53 \* CulDSac - 12788.64 \* TwnhsE - 20965.71 \* Twnhs - 42874.6 \* Duplex + 43318.45 \* KEx**

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + YearRemodel + TotalBsmtSF +
##     TotalBath + TotalSQFT + QualityIndex + TotRmsAbvGrd + GarageArea +
##     Tier1 + Tier2 + Tier3 + Tier4 + PartialSaleCond + CornerLot +
##     CulDSac + TwnhsE + Twnhs + Duplex + KEx, data = train.Clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -147708  -12715    -680   12197  200622
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.642e+05  9.522e+04  -2.774 0.005608 **
## LotArea       5.298e-01  9.467e-02   5.596 2.66e-08 ***
## YearRemodel   1.400e+02  4.845e+01   2.890 0.003912 **
## TotalBsmtSF  -1.393e+01  4.076e+00  -3.418 0.000649 ***
## TotalBath     8.798e+03  1.046e+03   8.408 < 2e-16 ***
## TotalSQFT     5.795e+01  3.069e+00  18.884 < 2e-16 ***
## QualityIndex  1.461e+03  1.005e+02  14.533 < 2e-16 ***
## TotRmsAbvGrd -1.819e+03  8.052e+02  -2.259 0.024048 *
## GarageArea    4.156e+01  4.622e+00   8.991 < 2e-16 ***
```

```
## Tier1          -6.430e+04  5.643e+03 -11.395 < 2e-16 ***
## Tier2          -5.889e+04  5.098e+03 -11.551 < 2e-16 ***
## Tier3          -4.709e+04  4.902e+03 -9.607 < 2e-16 ***
## Tier4          -2.420e+04  4.850e+03 -4.989 6.86e-07 ***
## PartialSaleCond 1.324e+04  2.497e+03  5.301 1.34e-07 ***
## CornerLot      -3.003e+03  1.800e+03 -1.669 0.095440 .
## CulDSac        7.598e+03  2.532e+03  3.001 0.002744 **
## TwnhsE        -1.279e+04  3.198e+03 -3.999 6.70e-05 ***
## Twnhs         -2.097e+04  5.437e+03 -3.856 0.000121 ***
## Duplex         -4.287e+04  4.462e+03 -9.609 < 2e-16 ***
## KEx           4.332e+04  2.900e+03 14.939 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24180 on 1347 degrees of freedom
## Multiple R-squared:  0.898, Adjusted R-squared:  0.8966
## F-statistic: 624.3 on 19 and 1347 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 = -264173.38 + 57.95 \* TotalSQFT + 1461.02 \* QualityIndex + 43318.45 \* KEx + 41.56 \* GarageArea - 58888.49 \* Tier2 - 42874.6 \* Duplex + 8797.95 \* TotalBath - 64301.63 \* Tier1 - 47087.15 \* Tier3 + 0.53 \* LotArea + 13237.91 \* PartialSaleCond - 24196.97 \* Tier4 - 12788.64 \* TwnhsE - 20965.71 \* Twnhs + 7597.53 \* CulDSac + 140.02 \* YearRemodel - 13.93 \* TotalBsmtSF - 1818.83 \* TotRmsAbvGrd - 3002.58 \* CornerLot**

```
##
## Call:
## lm(formula = SalePrice ~ TotalSQFT + QualityIndex + KEx + GarageArea +
##      Tier2 + Duplex + TotalBath + Tier1 + Tier3 + LotArea + PartialSaleCond +
##      Tier4 + TwnhsE + Twnhs + CulDSac + YearRemodel + TotalBsmtSF +
##      TotRmsAbvGrd + CornerLot, data = train.Clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -147708  -12715    -680   12197  200622
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.642e+05  9.522e+04  -2.774 0.005608 **
## TotalSQFT      5.795e+01  3.069e+00  18.884 < 2e-16 ***
## QualityIndex   1.461e+03  1.005e+02  14.533 < 2e-16 ***
## KEx           4.332e+04  2.900e+03  14.939 < 2e-16 ***
## GarageArea     4.156e+01  4.622e+00   8.991 < 2e-16 ***
## Tier2        -5.889e+04  5.098e+03 -11.551 < 2e-16 ***
## Duplex        -4.287e+04  4.462e+03 -9.609 < 2e-16 ***
## TotalBath      8.798e+03  1.046e+03   8.408 < 2e-16 ***
## Tier1        -6.430e+04  5.643e+03 -11.395 < 2e-16 ***
## Tier3        -4.709e+04  4.902e+03 -9.607 < 2e-16 ***
## LotArea        5.298e-01  9.467e-02   5.596 2.66e-08 ***
## PartialSaleCond 1.324e+04  2.497e+03  5.301 1.34e-07 ***
```

```
## Tier4          -2.420e+04  4.850e+03  -4.989  6.86e-07 ***
## TwnhsE         -1.279e+04  3.198e+03  -3.999  6.70e-05 ***
## Twnhs          -2.097e+04  5.437e+03  -3.856  0.000121 ***
## CulDSac        7.598e+03  2.532e+03   3.001  0.002744 **
## YearRemodel    1.400e+02  4.845e+01   2.890  0.003912 **
## TotalBsmtSF    -1.393e+01  4.076e+00  -3.418  0.000649 ***
## TotRmsAbvGrd   -1.819e+03  8.052e+02  -2.259  0.024048 *
## CornerLot      -3.003e+03  1.800e+03  -1.669  0.095440 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24180 on 1347 degrees of freedom
## Multiple R-squared:  0.898, Adjusted R-squared:  0.8966
## F-statistic: 624.3 on 19 and 1347 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 OverallCond. TotalSQFT is a result of a linear combination of variables that include GrLivArea. Below is the junk model and its fit.

**SalePrice = -258690.78 + 43765.66 \* OverallQual + 24815.4 \* OverallCond - 3350.1 \* QualityIndex + 17.77 \* GrLivArea + 48.29 \* TotalSQFT**

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + OverallCond + QualityIndex +
##     GrLivArea + TotalSQFT, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -139775 -18764    -445    16287   234441
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.587e+05  2.157e+04 -11.994 < 2e-16 ***
## OverallQual  4.377e+04  3.654e+03  11.976 < 2e-16 ***
## OverallCond  2.482e+04  4.014e+03   6.182 8.33e-10 ***
## QualityIndex -3.350e+03  6.931e+02  -4.833 1.50e-06 ***
## GrLivArea    1.777e+01  3.513e+00   5.057 4.83e-07 ***
## TotalSQFT    4.829e+01  2.498e+00  19.329 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30650 on 1361 degrees of freedom
## Multiple R-squared:  0.8345, Adjusted R-squared:  0.8339
## F-statistic: 1372 on 5 and 1361 DF, p-value: < 2.2e-16
```