Statistical Modeling and Analysis Results for Housing Prices in Ames
Kaggle Project
MSDS 6371

Submitted to:
Dr. Charles South

Report prepared by:
Tanvi Arora
Anjli Solsi
Olha Tanyuk

April 15, 2018

Introduction

This report summarizes the statistical modeling and analysis results for the residential housing data set in Ames, Iowa. Analysis is limited to the techniques learned in MSDS 6371. The purpose of this report is to document the detailed analysis of the proposed questions regarding the data set.
For Analysis Question 1, an explanation is made to determine the relationship between the sale prices and living areas of the homes, while limiting the sample to 3 specified neighborhoods from the dataset.

For Analysis Question 2, predictions on the sales prices of homes are made for the test dataset by creating models with known techniques from the training dataset. The variable selection techniques used are: forward selection, backward selection, and stepwise selection. These models are compared using the adjusted R^2 values and CV Press statistics.

The final section includes an Appendix with SAS code for each analysis question and screenshots with additional details.

Data Description

Data is provided from the Kaggle competition, “House Prices: Advanced Regression Techniques”. It is the housing dataset for Ames, Iowa, compiled by Dean De Cock for use in data science education.

With 79 explanatory variables, describing almost every aspect of a residential homes in Ames, Iowa; there is a training dataset with 1460 records and the resulting SalePrice. The test dataset of 1459 records are provided with the exact same 79 explanatory variables.

Files used:

Train.csv – the training dataset

TestCleaned.csv – cleaned test dataset

Analysis Question 1

Problem

Get an estimate of how the sale price (SalePrice) of the house is related to the square footage of the living area of the house (GrLIvArea) and if the SalePrice (and its relationship to square footage) depends on which neighborhood the house is located in. The company is only concerned with the three neighborhoods they sell in: NAmes, Edwards, and BrkSide. Realtors prefer to talk about living area in increments of 100 sq. ft. The variables used in this analysis are Neighborhood, GrLivArea, and SalePrice.

Data Transformation

Transform Living Area data (GrLIvArea) in order to get living area in increments of 100 sq. ft.

Build and Fit the Model

It is evident from the base model that the SalePrice of the house is definitely related to the square footage of the living area of the house (GrLIvArea_100sq_ft), and the SalePrice (and its relationship to square footage) depends on which neighborhood the house is located in: R-Square = 0.45. Variables in the model are significant, except one, Neighborhood Edwards (p=0.1336).

The model is:

$$\widehat{SalesPrice} = 74,676.40 + 5431.59 * GrLIvArea\_100sq\_ft - 54704.89 * Neighborhood\ BrkSide +$$
$$+ 13676.70 * Neighborhood\ Edwards + 3284.67 * GrLIvArea\_100sq\_ft * Neighborhood\ BrkSide +$$
$$- 2456.56 * GrLIvArea\_100sq\_ft * Neighborhood\ Edwards.$$

Assumptions

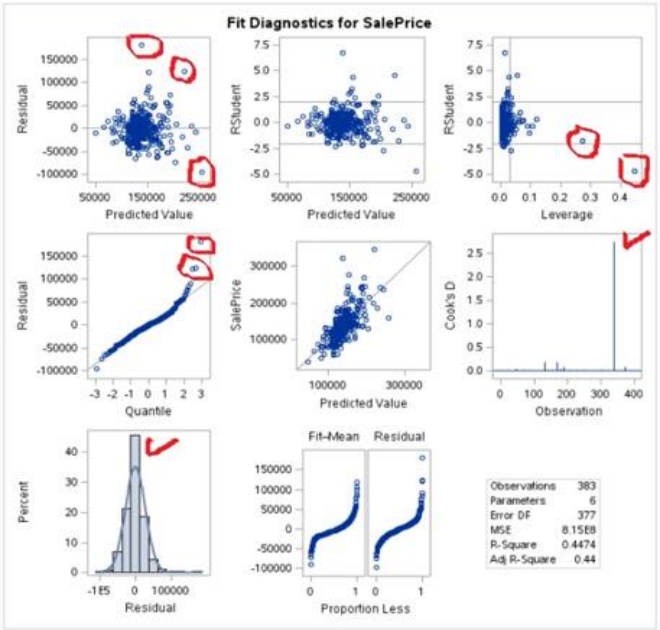

Figure 1.1 Fit Diagnostics for Original Data

1. Linearity: The line is almost straight with visually significant outliers at the right end on the QQ plot.

2. Constant Variance: The residual plot is a random cloud for most of the points. Most of the data is also within the range for studentized residuals. The RStudent plot for leverages shows two points that seem to be high. The Cook's D plot further supports this analysis with a high number (2.5). Centering can help resolve the collinearity issue.
3. Normality: The residual histogram presents a long tail towards the right, which may be due to the outliers.
4. Independence: The observations are assumed to be independent, but that cannot be applied to the variables with certainty. Independence will be assumed.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	80326	5592.03832	14.36	<.0001	0
GrLivArea_100sq_ft	GrLivArea_100sq.ft	1	4956.12477	409.70671	12.10	<.0001	1.58092
d1		1	-60354	12060	-5.00	<.0001	11.70999
d2		1	-43225	10838	-3.99	<.0001	13.89411
int1		1	3760.12849	940.21789	4.00	<.0001	11.53582
int2		1	2059.71212	820.38610	2.51	0.0125	13.92369

Figure 1.2 Variance Inflation Factor for Original Data

Data Transformation

Variance Inflation for neighborhoods and interaction variables looks higher than it should (>10). This is reduced by centering the living area variable. This does not impact the model.

After investigating the training dataset for high leverage points, the below records are removed, as they do not represent the majority of the samples. These specific houses in Edwards neighborhood are comparatively big for the sale price. There could be other variables like ages of the building or overall condition that could have affected the sale price. More data is required to analyze these houses.

Obs. 131: Neighborhood Edwards 46.76 100sq_ft $184,750
Obs. 339: Neighborhood Edwards 56.42 100sq_ft $160,000

The prices of these houses are unusually high, indicating that other variables may be involved beyond living area and neighborhood, such as modern renovations. After inspecting the potential outliers, it was determined that these houses are not representative of the sample of the houses in the neighborhoods being studied. Therefore, the data points were removed. More data is required to get a better understanding of these kinds of houses.

Obs. 169: Neighborhood NAmes 27.04 100sq_ft $345,000
Obs. 190: Neighborhood Edwards 16.98 100sq_ft $320,000

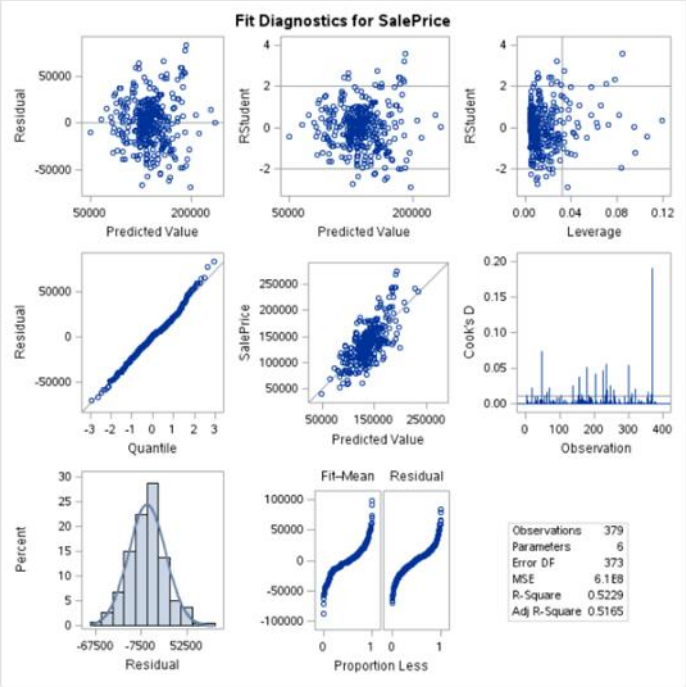

Figure 1.3 Fit Diagnostics for Data with Outliers removed

Assumptions for Transformed Data

1. Linearity: The QQ plots shows a very significant straight line with small outliers. Linearity can be assumed.
2. Constant Variance: There is a random spread of residuals, without outliers. The studentized residual plot shows visual significance, with approximately 95% of the data within the acceptable range of (-2, 2). The leverage plot shows an acceptable range of leverage points. The Cook's D plot looks better, having a maximum value of 0.2.
3. Normality: The residual histogram closely resembles a normal distribution with a slight right skew. Normality can be assumed.
4. Independence: The observations are assumed to be independent, but that cannot be applied to the variables with certainty. Independence will be assumed.

Since all assumptions are met, this model will be used for the analysis.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	66249	4595.17334	14.42	<.0001	0	57213	75285
GrLivArea_100sq_ft	GrLivArea_100sq.ft	1	6058.71073	329.39579	18.39	<.0001	1.02188	5411.00519	6706.41628
d1		1	-12349	3693.84677	-3.34	0.0009	1.09854	-19612	-5085.26359
d2		1	-16929	3007.49947	-5.63	<.0001	1.06994	-22843	-11015
cent1		1	3760.12849	940.21789	4.00	<.0001	1.07620	1911.33841	5608.92057
cent2		1	2059.71212	820.38610	2.51	0.0125	1.06647	446.55059	3672.87364

Figure 1.4 Variance Inflation Factor for Data with Outliers removed

Compare Competing Models

	Adjusted R^2	CV Press
No records removed	0.4400	3.42227
Outliers and leverage points removed	0.5165	2.4254

The resulting model is:
*SalePrice = 66,249 + 6,058.71*GrLivArea_100sq_ft -12,349*d1-16,929* d2 + 3,760.42 * cent1 + 2,059.71*cent2*
OR
*SalePrice = 80,329.59 + 4,956.12*GrLivArea_100sq_ft - 60,369.56*BrkSide - 43,231.5*Edwards + 3,760.12*GrLivArea_100sq_ft*BrkSide + 2,059.71*GrLivArea_100sq_ft*Edwards*

Parameters
The effect of the Living Area on Sale Price is 4,956.12 + 3,760.12**BrkSide + 2,059.71*Edwards.*
The effect of the Living Area on Sale Price when the neighborhood is BrkSide is 4,956.12 + 3,760.12 = $8,716.24.
The effect of the Living Area on Sale Price when the neighborhood is Edwards is 4,956.12 + 2,059.71 = $7,015.83.
The effect of the Living Area on Sale Price the when neighborhood is NAmes is $4,956.12.

Conclusion
We confirm that the size of the living area and specific neighborhood can explain 52.3% of variability of the house sale price.

Analysis Question 2

Problem
Given the data “train.csv” of 79 explanatory variables describing many aspects of residential homes in Ames, Iowa, the goal is to build the most predictive model for sales prices of homes in the dataset “test.csv”. Using the methods learned throughout the semester, four models are to be produced using the various selection methods learned and a custom model.

Model Selection
Stepwise
The variables determined by proc glmselect to be used for the stepwise selection model were: MSZoning, Neighborhood, OverallQual, OverallCond, YearBuilt, BsmtFinSF1, TotalBsmtSF, GrLivArea, Fireplaces, GarageArea, and SaleCondition. Please see Figure 2.7 for more details on the model.

Forward
The variables determined by proc glmselect to be used for the forward selection model were: MSZoning, LotArea, Neighborhood, OverallQual, OverallCond, YearBuilt, YearRemodAdd, BsmtFinSF1, TotalBsmtSF, CentralAir, GrLivArea, BsmtFullBath, KitchenAbvGr, KitchenQual, Functional, Fireplaces, GarageCars, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, ScreenPorch, and SaleCondition. Please see Figure 2.10 for more details on the model.

Backward
The variables determined by proc glmselect to be used for the backward selection model were: GarageFinish, GarageType, MSZoning, LotArea, Street, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, BldgType, OverallQual, OverallCond, YearBuilt, YearRemodAdd, RoofStyle, RoofMatl, MasVnrType, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtExposure, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, Heating, HeatingQC, CentralAir, Electrical, _1stFlrSF, _2ndFlrSF, LowQualFinSF, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr, KitchenQual,

TotRmsAbvGrd, Functional, Fireplaces, GarageYrBlt, GarageCars, GarageArea, GarageQual, GarageCond, PavedDrive, WoodDeckSF, OpenPorchSF, EnclosedPorch, _3SsnPorch, ScreenPorch, PoolArea, PoolQC, Fence, MiscFeature, MiscVal, YrSold, SaleType, and SaleCondition. Please see Figure 2.4 for more details on the model.

Custom
The group determined that the stepwise model would also be used for the custom model. This considered the most data with the best predictive values out of all the models.

Data Transformation

To deal with missing values, the training and test datasets have been modified to replace “NA” values with a blank. For categorical values, if they represent a feature of the house, we assumed NA represents “Not Available” and treated it as a category. Blanks in all category variables have been replaced with “NA”, such as in the original data.

The following are categorical variables for which NA was treated as a category: Alley, MasVnrType, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, Electrical, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, PoolQC, Fence, MiscFeature, and Utilities.

LotFrontage: The area of each street connected to the properties most likely has a similar area to other houses in the neighborhood. Since all neighborhoods are in the same city, the median of LotFrontage was assigned to the null values. A better way would have been to use the median of neighborhood, rather than the whole city.

masVnrArea: This is NA where masVnrType is NA. Assuming masVnrType is NA, an assumption is made that masonry veneering is not done; therefore, masVnrArea = 0 sq feet if it is NA.

GarageYrBlt: This is NA where other features of Garage are missing. When the other features are missing, an assumption is made that there is no garage, and GarageYrBlt = 0.

Although non-quantitative variables are mostly treated as categorical; this group assumed some of the quantitative variables with fixed values (like a group) and those thought to be significant as categorical in the models. Those variables are: OverallCond, MSSubClass, MoSold, and YrSold.

For data modifications, the Kaggle discussion board was used as a reference.

Assumptions

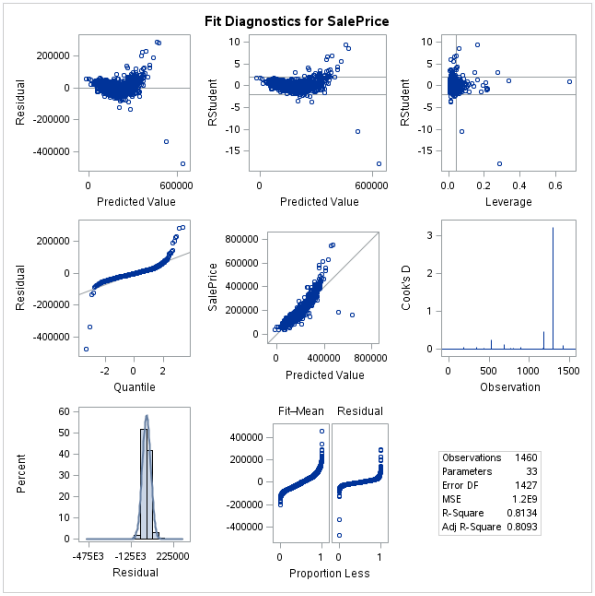

Figure 2.1 Fit Diagnostics for Original Data

1. Linearity: The line is almost straight with curved ends on the QQ plot. There are a few very significant outliers at both ends.
2. Constant Variance: Most of the data is clouded together with some visually distinct points. The studentized residuals have a higher range, and the leverage plot shows distinct points that have high leverage. Cook's D is 3, which is not acceptable.
3. Normality: There is a long tail on the residual histogram. This needs to be checked again with outliers deleted (if determined to do so).
4. Independence: The observations are assumed to be independent, but that cannot be applied to the variables with certainty. Independence will be assumed.

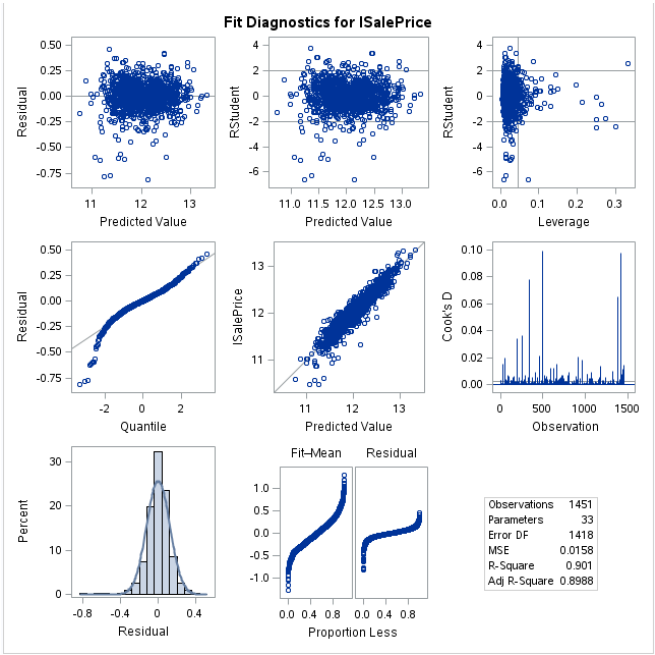

Figure 2.2 Fit Diagnostics for Log-Transformed data

The original data failed most of the required assumptions and a few changes were made to resolve this:

- The dependent variable was converted to a log scale to get better linearity.
- The id's identified below seemed to have high leverage points due to their distinct nature, which was not common amongst the houses in the sample. After inspecting the training dataset for potential outliers and leverage points, about 9 houses were found to not be representative of the samples in the dataset. It was determined that more data was required to analyze houses with such similar features. Id's of the houses excluded are: 524, 1299, 347, 314, 1231, 725, 643, 692, and 1183.

1. Linearity: The line is almost straight with curved ends on the QQ plot. There are a few outliers at both the ends with a smaller range, so linearity can be assumed on a log scale.
2. Constant Variance: The residual plot resembles a random cloud, so constant variance can be assumed. About 95% of the samples are within the range (2, -2) in the studentized residual plot. This is an acceptable limit. The leverage plot shows a few points that have some leverage (high leverage points are removed). Cook's D is 0.1, which is very good.
3. Normality: The histogram looks normally distributed, which is good.
4. Independence: The observations are assumed to be independent, but that cannot be applied to the variables with certainty. Independence will be assumed.

Compare Competing Models

Predictive Models	**Adjusted R^2**	**CV Press**	**Kaggle Score**
Forward	0.9263	18.60842	0.13870
Backward	0.9359	21.13804	0.15661
Stepwise	0.9128	21.15243	0.13567
Custom	0.9128	21.15243	0.13567

Conclusion

It is important to note that each model has its own advantages and disadvantages, with no model being the best for the data set. The highest adjusted R^2 value came from the backward selection model, and the lowest cross-validation value came from the forward selection model. Based on this dataset and analysis of models, we determined the forward selection model to be the most predictive for the sales prices of houses. In addition to the cross-validation score being the lowest value of 18.60842, this model provides the smallest root mean square error of 0.10729. We were surprised that stepwise selection provided the most predictive model, but the statistics support forward selection as the better model.

Appendix

Analysis 1

Figures/Tables

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.447376	20.68070	28552.30	138062.5

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	74676.40154	B	6337.89399	11.78	<.0001
GrLivArea_100sq_ft	5431.58627	B	461.36383	11.77	<.0001
Neighborhood BrkSide	-54704.88774	B	13882.33384	-3.94	<.0001
Neighborhood Edwards	13876.70324	B	9097.57465	1.50	0.1336
Neighborhood NAmes	0.00000	B	.	.	.
GrLivArea*Neighborho BrkSide	3284.66699	B	1081.53753	3.04	0.0026
GrLivArea*Neighborho Edwards	-2456.55603	B	636.13906	-3.86	0.0001
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.

Root MSE	28552
Dependent Mean	138063
R-Square	0.4474
Adj R-Sq	0.4400
AIC	8249.72388
AICC	8250.02255
SBC	7888.41209
CV PRESS	**3.422277E11**

Figure 1.5 Output Statistics for Original Data

Root MSE	24701	R-Square	0.5229
Dependent Mean	136855	Adj R-Sq	0.5165
Coeff Var	18.04909		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	80326	5592.03832	14.36	<.0001	69330	91322
GrLivArea_100sq_ft	GrLivArea_100sq.ft	1	4956.12477	409.70671	12.10	<.0001	4150.50033	5761.74922
d1		1	-60354	12060	-5.00	<.0001	-84068	-36640
d2		1	-43225	10838	-3.99	<.0001	-64536	-21914
int1		1	3760.12849	940.21789	4.00	<.0001	1911.33641	5608.92057
int2		1	2059.71212	820.38610	2.51	0.0125	446.55059	3672.87364

CV PRESS	**2.425404E11**

Figure 1.6 Output Statistics for Data with Outliers removed

SAS Code

```
proc glm data=train_Q1 plot=all;
class Neighborhood;
model SalePrice=GrLivArea_100sq_ft | Neighborhood / solution;
run;
```

```
*Delete outliers;
data train_Q1_2;
set train_Q1;
if _n_ = 339 then delete;
if _n_ = 131 then delete;
if _n_ = 169 then delete;
if _n_ = 190 then delete;
run;

data train_Q1_3;
set train_Q1_2;
if Neighborhood = "BrkSide" then d1=1; else d1=0;
if Neighborhood = "Edwards" then d2=1; else d2=0;
int1=d1*GrLivArea_100sq_ft;
int2=d2*GrLivArea_100sq_ft;
run;
proc reg data = train_Q1_3;
model SalePrice = GrLivArea_100sq_ft d1 d2 int1 int2 / clb ;
title "Regression of Sale Price with Interaction Terms";
run; quit;
proc glmselect data=train_Q1_3;
class Neighborhood;
model SalePrice = GrLivArea_100sq_ft | Neighborhood /
selection=backward(choose=CV stop=CV) cvmethod=split(10) CVdetails;
run;

*Check variance inflation factor (VIF) for variables in the model;
data train_Q1_3;
set train_Q1_2;
if Neighborhood = "BrkSide" then d1=1; else d1=0;
if Neighborhood = "Edwards" then d2=1; else d2=0;
int1=d1*GrLivArea_100sq_ft;
int2=d2*GrLivArea_100sq_ft;
run;
proc reg data = train_Q1_3;
model SalePrice = GrLivArea_100sq_ft d1 d2 int1 int2 / VIF ;
title "Regression of Sale Price with Interaction Terms";
run; quit;

*Variance Inflation for neighborhoods and interaction variables looks higher then it should (>10);
*Center interaction variables;
proc means data = train_Q1_3;
var GrLivArea_100sq_ft d1 d2;
run;
data center;
set train_Q1_3;
cent1 = (GrLivArea_100sq_ft-12.767)*(d1-0.153);
```

```
cent2 = (GrLivArea_100sq_ft-12.767)*(d2-0.256);
run;
proc reg data = center;
model SalePrice = GrLivArea_100sq_ft d1 d2 cent1 cent2 / VIF clb;
run;
```

Analysis 2

Figures/Tables

Root MSE	0.10009
Dependent Mean	12.02040
R-Square	0.9444
Adj R-Sq	**0.9359**
AIC	-5046.69466
AICC	-4985.78629
SBC	-5475.37306
CV PRESS	**21.13804**

Table 2.3 Results of Backward Selection Model

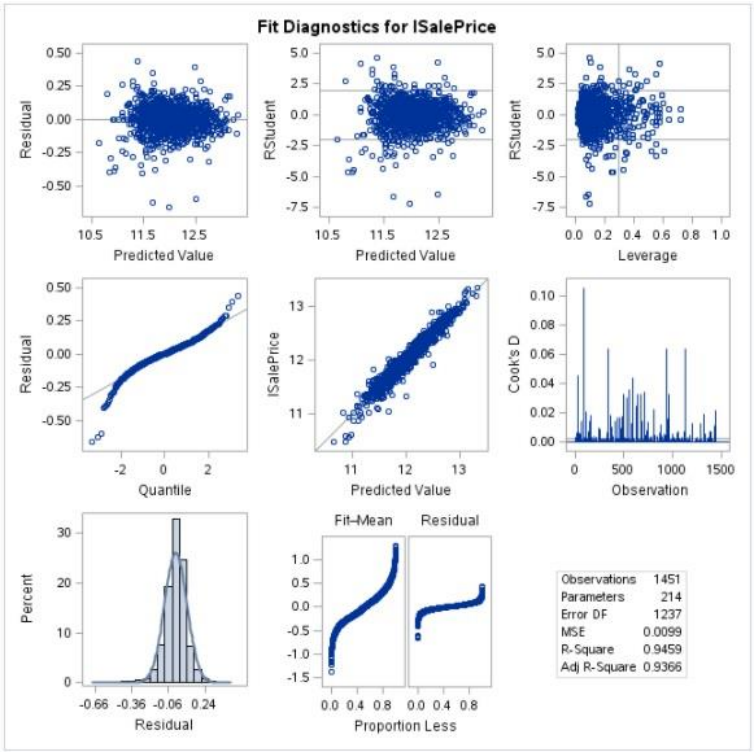

Figure 2.4 Fit Diagnostics for Backward Selection Model (Dependent Variable - log of SalePrice)

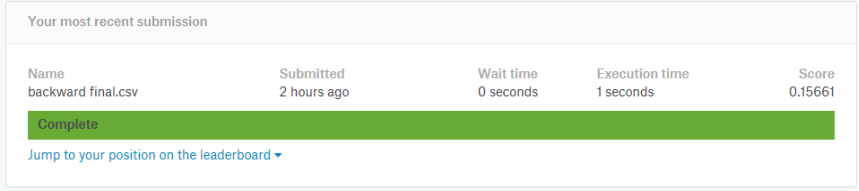

Figure 2.5 Kaggle Score for Backward Selection Model

Root MSE	0.11675
Dependent Mean	12.02040
R-Square	0.9157
Adj R-Sq	**0.9128**
AIC	-4731.45495
AICC	-4727.81209
SBC	-5925.73454
CV PRESS	**21.15243**

Table 2.6 Results of Stepwise Selection Model

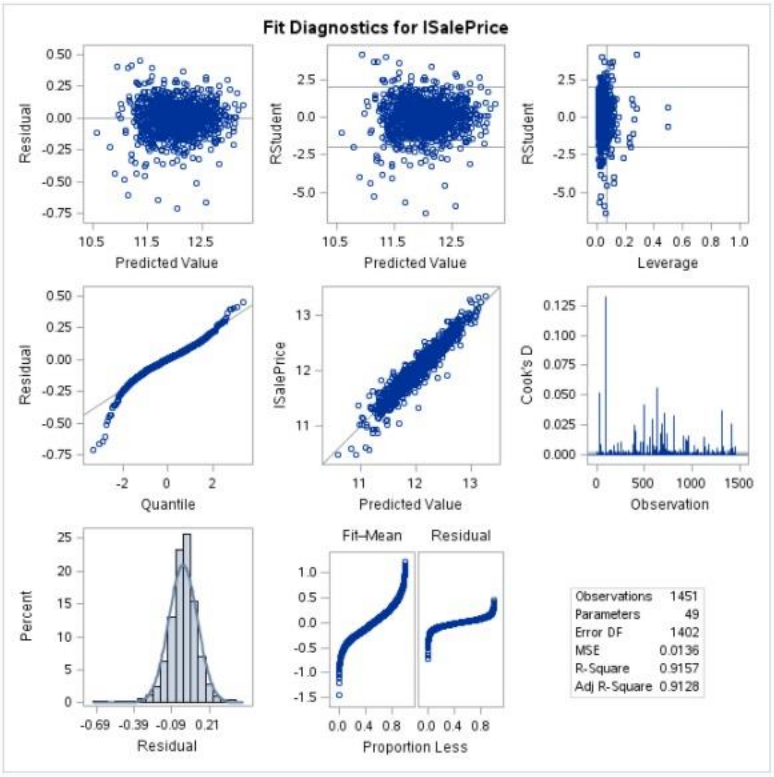

Figure 2.7 Fit Diagnostics for Stepwise Selection Model (Dependent Variable - log of SalePrice)

Figure 2.8 Kaggle Score for Stepwise Selection Model

Root MSE	0.10729
Dependent Mean	12.02040
R-Square	0.9297
Adj R-Sq	**0.9263**
AIC	-4958.57074
AICC	-4951.57581
SBC	-6052.53018
CV PRESS	**18.60842**

Table 2.9 Results of Forward Selection Model

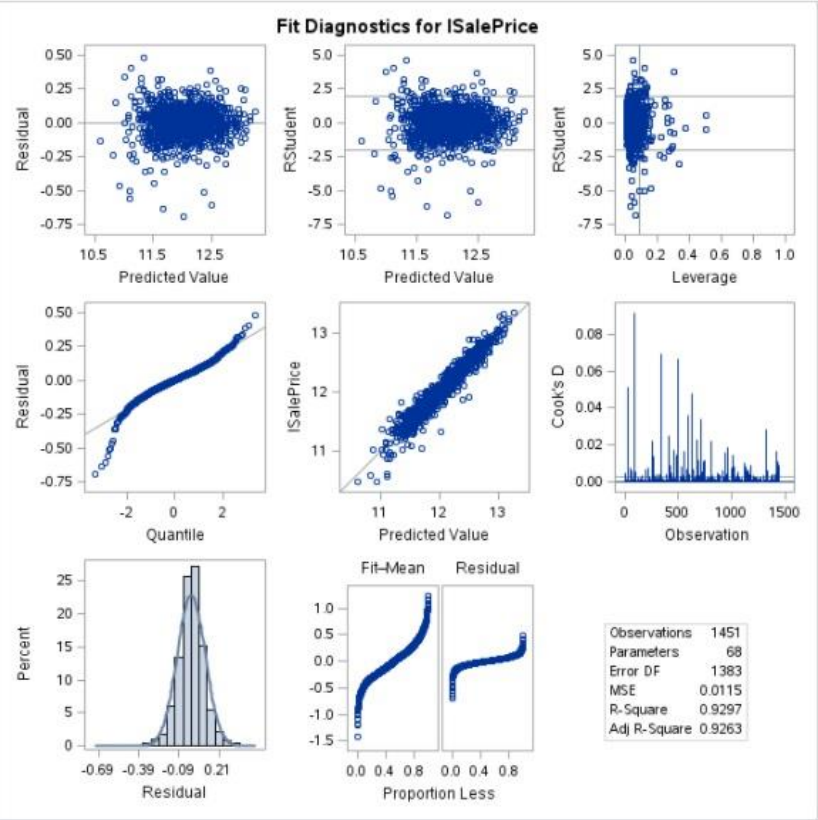

Figure 2.10 Fit Diagnostics for Forward Selection Model (Dependent Variable - log of SalePrice)

Figure 2.11 Kaggle Score for Forward Selection Model

SAS Code

**Kaggle Discussion board was referenced, and knowledge gained was applied to the code
NOTE: Source file is modified to replace NA with *blank* and loaded as an .xlsx file

```
/* Generated Code (IMPORT) */
/* Source File: trainCleaned.xlsx */
/* Source Path: /folders/myfolders/Project */
/* Code generated on: 4/14/18, 12:56 PM */

%web_drop_table(WORK.trainCleaned)

FILENAME REFFILE '/folders/myfolders/Project/trainCleaned.xlsx';

PROC IMPORT DATAFILE=REFFILE
   DBMS=XLSX
   OUT=WORK.trainCleaned;
   GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.trainCleaned; RUN;

/* Generated Code (IMPORT) */
/* Source File: testCleaned_noNA.xlsx */
/* Source Path: /folders/myfolders/Project */
/* Code generated on: 4/14/18, 12:56 PM */

%web_drop_table(WORK.testCleaned_noNA);

FILENAME REFFILE '/folders/myfolders/Project/testCleaned_noNA.xlsx';

PROC IMPORT DATAFILE=REFFILE
   DBMS=XLSX
   OUT=WORK.testCleaned_noNA;
   GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.testCleaned_noNA; RUN;
%web_open_table(WORK.testCleaned_noNA);

*Data transformation for the train data set;
*Remove leverage points and outliers;
data houseprice_train;
set trainCleaned;
if Alley Eq '' then Alley='NA' ;
if MasVnrType Eq '' then MasVnrType='NA' ;
if BsmtQual Eq '' then BsmtQual='NA' ;
if BsmtCond Eq '' then BsmtCond='NA' ;
if BsmtExposure Eq '' then BsmtExposure='NA' ;
if BsmtFinType1 Eq '' then BsmtFinType1='NA' ;
if BsmtFinType2 Eq '' then BsmtFinType2='NA' ;
if Electrical Eq '' then Electrical='NA' ;
if FireplaceQu Eq '' then FireplaceQu='NA' ;
```

```
if GarageType Eq '' then GarageType='NA' ;
if GarageFinish Eq '' then GarageFinish='NA' ;
if GarageQual Eq '' then GarageQual='NA' ;
if GarageCond Eq '' then GarageCond='NA' ;
if GarageYrBlt EQ '' then GarageYrBlt=0;
if PoolQC Eq '' then PoolQC='NA' ;
if Fence Eq '' then Fence='NA' ;
if MiscFeature Eq '' then MiscFeature='NA' ;
if Utilities EQ ''  then Utilities='NA';
*removing leverage;
if Id EQ 524 then delete;
if ID EQ 1299  then delete;
if Id EQ 347  then delete;
if ID EQ 314 then delete;
if ID EQ 1231 then delete;
* removing outliers;
if Id EQ 725 OR Id EQ 643 then delete;
if ID EQ 692 OR ID EQ 1183 then delete;
lSalePrice = log(SalePrice);
if notdigit(lotFrontage) then lotFrontage = 69;
if notdigit(masVnrArea) then  masVnrArea = 0;
run;

proc reg data=houseprice_train;
model lSalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1
BsmtFinSF2 BsmtUnfSF TotalBsmtSF OstFlrSF TndFlrSF LowQualFinSF GrLivArea BsmtFullBath
BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt
GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch TSsnPorch ScreenPorch PoolArea
MiscVal MoSold YrSold / VIF;
run;

*Data transformation for the test data set;
data houseprice_test;
set testCleaned_noNA;
if Alley Eq '' then Alley='NA' ;
if MasVnrType Eq '' then MasVnrType='NA' ;
if BsmtQual Eq '' then BsmtQual='NA' ;
if BsmtCond Eq '' then BsmtCond='NA' ;
if BsmtExposure Eq '' then BsmtExposure='NA' ;
if BsmtFinType1 Eq '' then BsmtFinType1='NA' ;
if BsmtFinType2 Eq '' then BsmtFinType2='NA' ;
if Electrical Eq '' then Electrical='NA' ;
if FireplaceQu Eq '' then FireplaceQu='NA' ;
if GarageType Eq '' then GarageType='NA' ;
if GarageFinish Eq '' then GarageFinish='NA' ;
if GarageQual Eq '' then GarageQual='NA' ;
if GarageCond Eq '' then GarageCond='NA' ;
if GarageYrBlt EQ '' then GarageYrBlt=0;
```

```
if PoolQC Eq '' then PoolQC='NA' ;
if Fence Eq '' then Fence='NA' ;
if MiscFeature Eq '' then MiscFeature='NA' ;
if Utilities EQ ''  then Utilities='NA';
if notdigit(lotFrontage) then lotFrontage = 67;
if notdigit(masVnrArea) then  masVnrArea = 0;
SalePrice = .;
lSalePrice = .;
run;

*Combined data set;
data home_estate;
set houseprice_Train houseprice_test;
run;

*Run proc glmselect to determine variables used for each selection model;
*Note: the other models are commented out at the end;
proc glmselect   data=houseprice_train seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot
CRITERIONPANEL);
class  OverallCond MSSubClass MoSold YrSold GarageFinish GarageType MSZoning Street LotShape
LandContour
Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle
RoofMatl
Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond
BsmtExposure BsmtFinType1
BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageQual
GarageCond
PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;
model lSalePrice =  GarageFinish GarageType MSSubClass LotFrontage MasVnrArea MSZoning LotArea
Street
LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
HouseStyle
OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd
MasVnrType ExterQual
ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
BsmtFinSF2 BsmtUnfSF
TotalBsmtSF Heating HeatingQC CentralAir Electrical _1stFlrSF _2ndFlrSF LowQualFinSF GrLivArea
BsmtFullBath
BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
Fireplaces
FireplaceQu GarageYrBlt GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF
OpenPorchSF
EnclosedPorch _3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature MiscVal MoSold YrSold
SaleType
SaleCondition /
selection = backward (choose=CV stop=CV) cvmethod=split(10) CVdetails;
*selection = forward (choose=CV stop=CV) cvmethod=split(10) CVdetails;
*selection = stepwise (choose=CV stop=CV) cvmethod=split(10) CVdetails;
```

```
run;

*Stepwise selection model;
proc glm data = home_estate plots = all;
class MSZoning  Neighborhood OverallCond SaleCondition;
model LSalePrice = MSZoning Neighborhood OverallQual OverallCond YearBuilt BsmtFinSF1 TotalBsmtSF
GrLivArea Fireplaces GarageArea SaleCondition
/ cli solution;
output out = results p = Predict;
run;

*Forward selection model;
proc glm data = home_estate plots = all;
class MSZoning  Neighborhood OverallCond SaleCondition CentralAir KitchenQual Functional
SaleCondition;
model LSalePrice = MSZoning LotArea Neighborhood OverallQual OverallCond YearBuilt YearRemodAdd
BsmtFinSF1 TotalBsmtSF CentralAir GrLivArea BsmtFullBath KitchenAbvGr KitchenQual Functional
Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch
SaleCondition
/ cli solution;
output out = results p = Predict;
run;

*Backward selection model;
proc glm data = home_estate plots = all;
class GarageFinish GarageType MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
Neighborhood Condition1 BldgType OverallCond RoofStyle RoofMatl MasVnrType ExterQual ExterCond
Foundation
BsmtQual BsmtCond BsmtExposure Heating HeatingQC CentralAir Electrical KitchenQual Functional
GarageQual
GarageCond PavedDrive PoolArea PoolQC Fence MiscFeature MiscVal YrSold SaleType SaleCondition;
model lSalePrice = GarageFinish GarageType MSZoning LotArea Street LotShape LandContour Utilities
LotConfig LandSlope Neighborhood Condition1 BldgType OverallQual OverallCond YearBuilt
YearRemodAdd
RoofStyle RoofMatl MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
BsmtFinSF1
BsmtFinSF2 BsmtUnfSF Heating HeatingQC CentralAir Electrical _1stFlrSF _2ndFlrSF LowQualFinSF
BsmtFullBath
BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
Fireplaces
GarageYrBlt GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
EnclosedPorch
_3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature MiscVal YrSold SaleType SaleCondition
/ cli solution;
output out = results p = Predict;
run;

*Custom model, we opted to use stepwise selection;
```

```
proc glm data = home_estate plots = all;
class MSZoning  Neighborhood OverallCond SaleCondition;
model LSalePrice = MSZoning Neighborhood OverallQual OverallCond YearBuilt BsmtFinSF1 TotalBsmtSF
GrLivArea Fireplaces GarageArea SaleCondition
/ cli solution;
output out = results p = Predict;
run;

*Calculate back-transform of predicted values of SalePrice for test.csv;
*For prediction values less than zero, we chose to use the mean value;
data results;
set results;
if Predict > 0 then SalePrice=exp(Predict);
else if Predict < 0 then SalePrice=166109;
else if notdigit(SalePrice) then SalePrice=166109;
keep Id SalePrice;
where Id > 1460;
run;

*Print predicted SalesPrices;
proc print data=results;
run;
```