Predictive Model for House Sale Price in Ames, Iowa: Using Data from 2006 to 2010

Manisha Pednekar

Abstract

Given the contemporary housing Ames data set[1], this study investigated predictive models to satisfactorily address Century 21 Ames' questions of interest. Specifically, we developed: 1) simplified model to capture relationship between living area and sale price of the house, along with dependence on specific neighborhood, in three neighborhoods (NAmes, Edwards and BrkSide) in Ames, Iowa; 2) best possible predictive model using standard procedures and custom procedure to predict the sale price of individual house in all of the Ames Iowa. Our study indicates that house price increases linearly with living area (NAmes: $5770, Edwards: $4532, and BrkSide: $9120 per 100 sqft, $p<0.001$) in each of the neighborhoods where Century 21 Ames sells houses, with BrkSide neighborhood showing 58% ($p<0.002$) steeper increase per square feet than NAmes neighborhood. Kaggle score of 0.24 (0.736 R^2) for our custom procedure was better, compared to standard procedures (Forward , Backward , Stepwise). Our predictive model indicates that living area, age of the house, basement size, and garage area play prominent role in final sale price of the house.

Introduction

The Ames data set describes the sale of individual residential property in Ames, Iowa from 2006 to 2010[(1)]. With the modified version of dataset as provided by Kaggle for website competition [(2)], this project aims at showing both the graphical and numerical statistical characteristics of influential factors in sale price prediction for Century 21 Ames who sells houses only in the NAmes, Edwards and BrkSide neighborhoods. Specifically, two questions of interest are addressed here: (1) demonstrate the relationship between greater living area of the house (GrLivArea) and sale price in these three neighborhoods; (2) build the most predictive model for sales prices of homes in all neighborhoods in Ames Iowa area.

Data Analysis

There are two subsets for the Kaggle housing data: *train* data (ID 1-1460 with sale price) and *test* data (ID 1461-2919 where sale price is not provided). Both data contain 79 explanatory variables (23 nominal, 23 ordinals, 14 discrete, and 19 continuous) describing (almost) every aspect of residential homes in Ames, Iowa. The combined *train* and *test* data were used to study the relationship of the different variable on the final sale price, in order to select most prominent influential factors in the construction of the model. Explanatory variables were used in 1) forward selection (FS), 2) backward elimination (BE), 3) Stepwise selection (SW), and 4) LASSO selection procedure. With careful observation of returned variables from these procedures and visual inspection of scatterplots with data transformations we built our custom set of influential variables. Returned models were judged based on the adjusted R2, CV

[1] Dean De Cock, Journal of Statistics Education 2011. 19(3)

[2] www.kaggle.com/c/house-prices-advanced-regression-techniques

Press to fine tune our custom model using the *train* data. The combined *train* and *test* data were used for predicating the final sale price. Final individual house sale prices predicted from the resultant models from FS, BE, SW, and our own CUSTOM model were submitted to Kaggle competition.

First Question of interest: How the SalePrice of the house is related to the square footage of the living area of the house (GrLIvArea) and if it depends on which neighborhood (NAmes, Edwards and BrkSide) in Ames, Iowa, the house is located in.

Model Selection

Step 1: Visual Inspection

After initial inspection, there are 383 sales for NAmes, Edwards and BrkSide neighborhoods in Ames, Iowa. The general scatterplot (Figure 1A) shows the normal data distribution of sale price and GriLivArea with positive correlation, and the coded scatterplot (Figure 1B) displays the monotonic linear relationship for 3 neighborhoods (Names in blue color, Edwards in red color, BrkSide in black color). On additional note, transforming both X and Y to new scales (i.e. log-log transformation) did not make the straight-line assumption more defensible (data not shown).

Validity Check:

Assumptions for multiple linear regression model were checked using histograms and Q-Q plot analysis of residuals to ensure the validity of the model. Residuals after linear fit were inspected for extreme observations using Cook's D influence point and leverage analysis.

Simple linear regression model (SalePrice= $\beta_0+\beta_1$*GrLivArea) treating all neighborhood types as same was run on Normal sale data to eliminate confounding factor from other sale conditions.

Regression of SalePrice on GrLivArea treating all neighborhoods same

The GLM Procedure

Number of Observations Read	334
Number of Observations Used	334

Regression of SalePrice on GrLivArea treating all neighborhoods same

The GLM Procedure
Dependent Variable: SalePrice

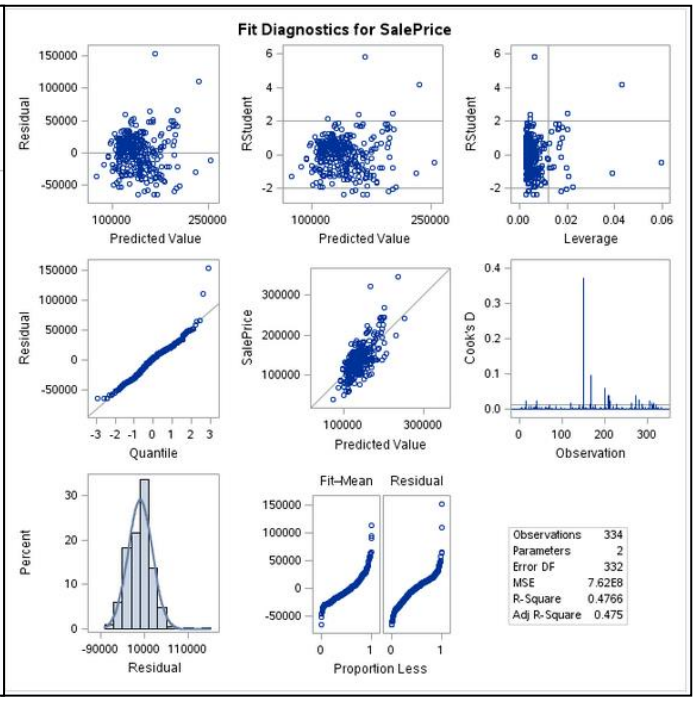
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	230459227004	230459227004	302.30	<.0001
Error	332	253102456497	762356796.68		
Corrected Total	333	483561683501			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.476587	19.84389	27610.81	139141.5

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GrLivArea	1	230459227004	230459227004	302.30	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GrLivArea	1	230459227004	230459227004	302.30	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	52370.65528	5214.304093	10.04	<.0001	42113.41490	62627.89566
GrLivArea	67.53053	3.884027	17.39	<.0001	59.89012	75.17093



F-test based p<.0001 and p<.0001 for intercept and slope parameter along with R2 of 0.48 suggest linear model is reasonable.

Assumptions check with residuals

Normality: Judging from scatter plot, QQ plot and histogram of residuals there is some (left skew) but not strong evidence against normality.

Linear Trend: checking pairwise scatter plots coded by neighborhood indicates a strong linear trend between each SalePrice and GrLivArea.

Equal SD: There is little evidence of heteroscedasticity from the scatter plots of residuals.

Independence: We will assume the observations are independent.

There is no indication for necessity for data transformation.

Accounting for neighborhood types

Multiple linear regression model accounting for neighborhood types was run after eliminating extreme observations based on GrLivArea and SalePrice to avoid high influence points (additional comments in code).

Parallel lines, assuming same rate of change in mean housing price for change in square fit area. Model: SalePrice= $\beta_0+\beta_1$*GrLivArea +β_2* Edwards + β_3* BrkSide with Names as a reference neightborhood.

Different slopes, allowing different rate of change in mean housing price for change in square fit area in different neighborhoods.
Model: SalePrice= $\beta_0+\beta_1$*(GrLivArea) +β_2* Edwards + β_3* BrkSide + β_4* (GrLivArea) * Edwards + β_5* (GrLivArea) * BrkSide with Names as a reference neightborhood.

Results

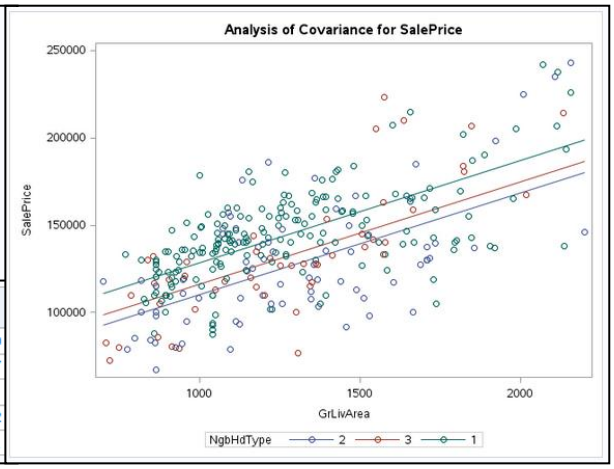
Regression of SalePrice on GrLivArea ignoring neighborhood type

The GLM Procedure
Dependent Variable: SalePrice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	137494511395	45831503798	84.90	<.0001
Error	312	168431283702	539843858.02		
Corrected Total	315	305925795097			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.449437	16.82094	23234.54	138128.7

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	70289.71634	B	5190.414994	13.54	<.0001	60077.07398	80502.35870
GrLivArea	58.34821		3.893108	14.99	<.0001	50.68814	66.00827
NgbHdType 2	-18497.83918	B	3143.056458	-5.89	<.0001	-24682.10606	-12313.57231
NgbHdType 3	-12218.96220	B	3722.825384	-3.28	0.0011	-19543.98039	-4893.94402
NgbHdType 1	0.00000	B	.	.	.	.	.



Regression of SalePrice on GrLivArea accounting for neighborhood type /w centering

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

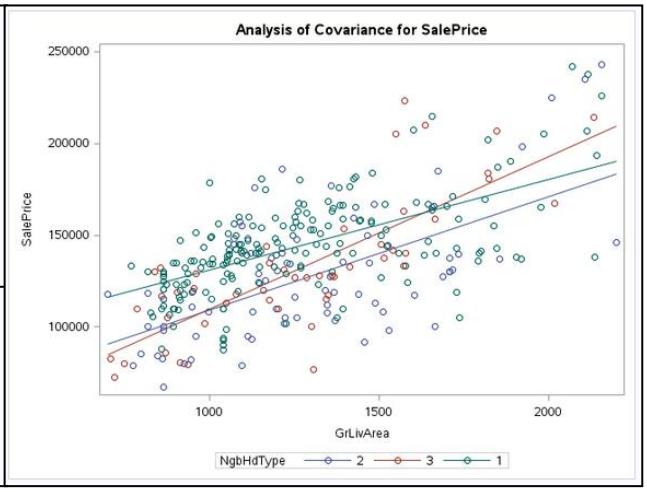
Number of Observations Read	316
Number of Observations Used	316

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1.428568E11	28571367658	54.32	<.0001
Error	310	1.63069E11	526028893		
Corrected Total	315	3.059258E11			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	71016	5129.85292	13.84	<.0001
GrLivArea	1	57.69810	3.84979	14.99	<.0001
d1	1	-18504	3107.45400	-5.95	<.0001
d2	1	-12029	3675.37168	-3.27	0.0012
cent1	1	12.37996	9.09685	1.36	0.1745
cent2	1	33.50539	10.64845	3.15	0.0018



Both the models provide F-test based p<0.001 and R^2 values are similar (0.45 and 0.47). However, separate slope model shows that rate of change of mean house price changes significantly faster (p<0.002) for BrkSide neighborhood than the other two neighborhoods.

Conclusion for the first QOI: In general, the mean individual house price increases by $6753 per 100sqft of greater living area in neighborhoods (NAmes, Edwards and BrkSide) where Century 21 Ames sells houses (two-side p value < 0.0001, 95% CI of 5989 to 7517). Specifically, rate of sale price increase in individual neighborhoods is: NAmes: $5770 (95% CI 50151,6524), Edwards: (95% CI 2749,6315), and BrkSide: $9120 (95% CI 7033,11208), per 100 sqft), with BrkSide neighborhood showing 58% (p<0.002) steeper increase per square feet than NAmes neighborhood.

Appendix A: SAS code for QOI1

```
/* Import data sets for sale prices of residential homes in Ames, Iowa */
/* with aspects of homes that potentially expain and help predict sale price */
/*Import training data set train.csv */
proc import datafile='/home/mpednekar0/BridgeCourse/kaggle_train.csv'
out = train DBMS=csv REPLACE; GETNAMES=YES;
run;

/* Pop out result in new browser tab for reference as we work along */
/*proc print data = train;
run; */


/* Input data file test.csv without sale price data*/
proc import datafile='/home/mpednekar0/BridgeCourse/testCleaned.csv'
out = test DBMS=csv REPLACE; GETNAMES=YES;
run;

/* Pop out result in new browser tab for reference as we work along */
/* proc print data = test;
run; */


/* Specific Aim 1 */

/* Century 21 Ames only sells houses in the NAmes, Edwards and BrkSide neighborhoods */
data trainNEBneighb;
  set train;
  if neighborhood in ('NAmes','Edwards','BrkSide');
run;

/* Century 21 would like to SIMPLY get an estimate of how the SalePrice of the house is related
to the square footage
 of the living area of the house (GrLIvArea) and if it depends on which neighborhood the house
is located in.  */

/* For simplified model lets use only Normal sale data to eliminate confounding factor from
other sale conditions */
Data trainNEBneighb_Norm;
set trainNEBneighb;
If SaleCondition ^= 'Normal' then delete;
run;
```

```
/* Quick check for linearity */
proc glm data=trainNEBneighb_Norm plots=all;
model SalePrice=GrLivArea /solution clparm;
title 'Regression of SalePrice on GrLivArea treating all neighborhoods same';
run;


/* Inspect scatter plot of sale price v/s GrLIvArea for reduced data size*/
proc sgscatter data = trainNEBneighb_Norm;
matrix SalePrice GrLivArea;
run;

/* Data set for these neighborhoods seems sparse for GrLIvArea > 2000 and SalePrice > 200K,
lets check boxplot, histogram and QQ plots */
proc univariate data = trainNEBneighb_Norm;
var GrLivArea;
histogram GrLivArea;
qqplot GrLivArea;
run;

proc univariate data = trainNEBneighb_Norm;
var SalePrice;
histogram SalePrice;
qqplot SalePrice;
run;


/* For simplified model lets eliminate Extreme Observations */
Data trainNEBneighb_Norm_Red;
set trainNEBneighb_Norm;
If GrLivArea >2200 or GrLivArea < 675 then delete;
If SalePrice >243900 or SalePrice < 61100 then delete;
/*
* Keep between 5 and 95 percentile
If GrLivArea >2108 or GrLivArea < 816 then delete;
If SalePrice >207500 or SalePrice < 82500 then delete;

* Keep between 10 and 90 percentile
If GrLivArea >1820 or GrLivArea < 864 then delete;
If SalePrice >180500 or SalePrice < 98000 then delete;
*/
run;

/* check if log transformation may help as both SalePrice and GrLivArea are right skewed */
```

```
/*
Data trainNEBneighb_Norm_lRed;
set trainNEBneighb_Norm_Red;
lSalePrice = log(SalePrice);
lGrLivArea = log(GrLivArea);
run; */
/* There was not significant improvement with log on linear regeression */
/* Plus for simplicity of inference we stick with original data */


/* Add neighborhood identifier to visualize neighboorhood factor in SalePrice */
data trainNEBneighb_Norm_Red_NType;
set trainNEBneighb_Norm_Red;
if Neighborhood = 'NAmes' then NgbHdType = 1;
  else if Neighborhood = 'Edwards' then NgbHdType = 2;
  else if Neighborhood = 'BrkSide' then NgbHdType = 3;
if Neighborhood = 'NAmes' then NgbHd = 'N';
  else if Neighborhood = 'Edwards' then NgbHd ='E';
  else if Neighborhood = 'BrkSide' then NgbHd ='B';
run;

symbol1 V=squarefilled C=black I=none;
symbol2 V=trianglefilled C=red I=none;
symbol3 V=diamondfilled C=blue I=none;
Title 'House Sales in NAmes, Edwards and BrkSide Neighborhoods from 2006 to 2010';
legend1 position=(right middle)
      label=(position=top)
      across=1;

/* Inspect scatter plot of sale price v/s GrLIvArea for reduced data size /w indicator for
neighborhoods */
proc gplot data=trainNEBneighb_Norm_Red_NType;
plot SalePrice * GrLivArea = NgbHd;
legend=legend1;
run;

data trainNEBneighb_Norm_Red_NInd;
set trainNEBneighb_Norm_Red_NType;
* coding indicator variable for neighborhood type;
if NgbHdType = 2 then d1 = 1; else d1 = 0;
if NgbHdType = 3 then d2 = 1; else d2 = 0;
int1 = d1*GrLivArea; int2 = d2*GrLivArea;
run;
```

```
/* Check equal slopes model first */
proc reg data=trainNEBneighb_Norm_Red_NInd;
model SalePrice = GrLivArea d1 d2;
title 'Regression of SalePrice on GrLivArea ignoring neighborhood type';
run; quit;

proc glm data=trainNEBneighb_Norm_Red_NType plots=all;
class NgbHdType (ref='1');
model SalePrice=GrLivArea NgbHdType/solution clparm;
title 'Regression of SalePrice on GrLivArea ignoring neighborhood type';
run;


/* Incorporate interaction between neighborhood and GrLivArea w/o centering the data */

proc reg data=trainNEBneighb_Norm_Red_NInd;
model SalePrice = GrLivArea d1 d2 int1 int2;
title 'Regression of SalePrice on GrLivArea accounting for neighborhood type';
run; quit;

proc glm data=trainNEBneighb_Norm_Red_NType plots=all;
class NgbHdType (ref='1');
model SalePrice=GrLivArea | NgbHdType/solution clparm;
title 'Regression of SalePrice on GrLivArea accounting for neighborhood type';
run;


/* Center the data */
proc means data=trainNEBneighb_Norm_Red_NInd;
var GrLivArea d1 d2;
run;
/*
GrLivArea 1272.38
d1 0.2436709
d2 0.1550633
*/

data centered;
set trainNEBneighb_Norm_Red_NInd;
cent1 = (GrLivArea - 1272.38)*(d1 - 0.2436709);
cent2 = (GrLivArea - 1272.38)*(d2 - 0.1550633);
run;

/* Incorporate interaction between neighborhood and GrLivArea w centered data */
```

```
proc reg data=centered;
model SalePrice = GrLivArea d1 d2 cent1 cent2;
title 'Regression of SalePrice on GrLivArea accounting for neighborhood type /w centering';
run; quit;

proc glm data=centered plots=all;
class NgbHdType (ref='1');
model SalePrice=GrLivArea | NgbHdType/solution clparm;
title 'Regression of SalePrice on GrLivArea accounting for neighborhood type  /w centering';
run;
```