ames iowa kaggle home price modeling

preeti swaminathan & patrick mcdevitt

14 avril 2017

House Prices - à la Kaggle - Homework 13

Analysis Question 1

Restatement of Problem:

Century 21 in Ames Iowa wants to build a model to predict sales price of 3 neighbourhood (BrkSide, NAmes, EdWard) in Iowa based on living area. They have provided with historical data of sales done in these neighborhood so far.
To-do: Build a model which uses (independent variables) living area in sq.ft and the neighborhood and predicts sales price (Dependent variable)

Analysis Question 2

Statement of Problem

Estimating market value of a home for sale has significant implications for all parties involved in the transaction : seller, buyer, agents, mortgage providers and even local taxing authorities. Getting it right can improve local economies. Inefficiencies associated to historical methods of value assessments create hesitation on the part of buyers and lenders, and potential loss of revenue for sellers and agents. Developing a model that considers all available factors and provides a transparent valuation that can be shared among all parties in the transaction can enable the participants to proceed with increased confidence, thus increasing the velocity of the local real estate market.
That is the purpose of this evaluation : use all available contributing factors for the residnetial real estate market in Ames, Iowa and create a predicitve model to better estimate market valuation for future properties to be proposed for sale.

Data Available & Utilized

For this evaluation, there are seventy-nine explanatory variables available for exploitation, based on residential sales in the years 2006 trough 2010, comprising approximately 1500 sales. The explanatory

variables include traditional expected characteristics, such as : neighborhood, square footage, number of bedrooms, number of bathrooms, etc. and also several factors that are perhaps considered secondary or tertiary, but are included in the modeling to increase predictive capability. Some of these additional factors include : heating type, number of fireplaces, qualitative assessment of the kitchen condition.

Model Construction

In order to build the model, the following steps are taken :
* read in the raw training data set provided,
* basic cleaning of the data, including removing significant outliers (for this purpose, more than 5 std deviations from mean)
* imputing values for features where none was provided (for this purpose, setting to mean value for numeric features, and creating a new factor level “None” for categorical features),
* plot and visually examine each feature in relation to log(SalePrice) …
+ this provides a basis for removing some features from consideration based on inspection
+ some features may have 1400 / 1460 within same cateegory, thus not providing varaibility worthwhile including in a model
+ some numerical features are sparsely populated, and the the few values visually exhibit zero slope in relation to log(SalePrice)
+ a new feature was created “saledate” from the “year sold” and “month sold” features. Upon visual examination, there was no obvious trend in the time series view for log(SalePrice)s, so this was eventually discarded (surprisingly, considering that the time period spanned the economic downturn to 2007 - 2009).
+ this visual examination then results in eliminating approximately 25 of the features from consideration in the model.
+ (All of the plots are available for review in the appendix, “homes_train_plots.pdf”, if desired for review)

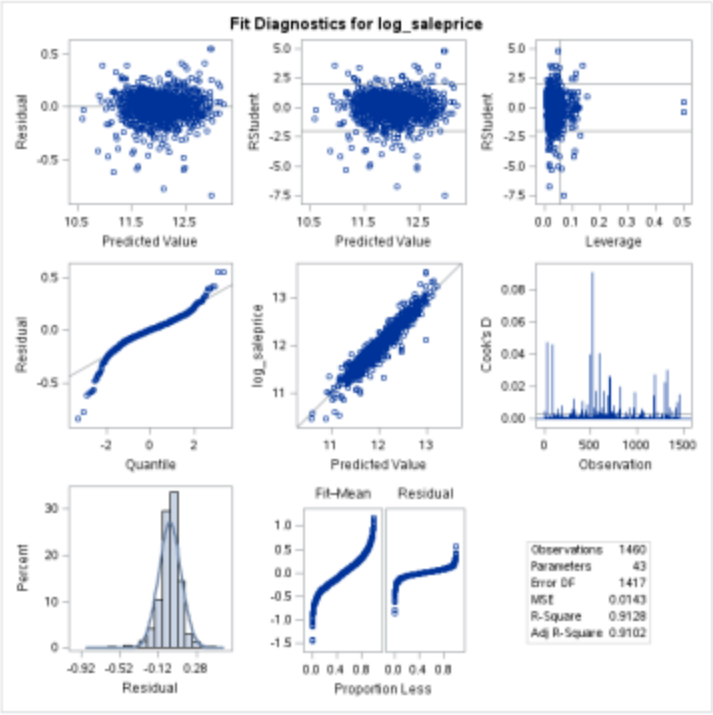
Models Considered

In all cases, the basic data set consists of 52 predictor variables and the dependent output variable log(SalePrice)

Four different models are built :
* Stepwise - modeled in SAS proc glmselect
* Forward - modeled in SAS proc glmselect
* Backward - modeled in SAS proc glmselect
* CUSTOM - model based on the above three models … just average the resutls of these models and see if this improves the Kaggle score.

Residuals Evaluation
* the following plots show the distribution of residuals and the predicted vs. the actual sale price (log

scale) for the custom model. From these plots, we consider that the basic assumptions of normality in the residuals is achieved, and that the visual aspect of the fitted model aligns well with the dependent variable to be modeled. For brevity's sake, the only plots shown here are for the stepwise selction model. The residuals plots for the other models assessed were similar in nature.



Residual Plots - SAS Stepwsie Model Results

Influential point analysis (Cook's D and Leverage)

Test for assumptions:

- Normality: From QQ plot, scatter and histogram, normality is generally respected. The residuals histogram and QQ plot show a slight tendency towards left skewness. With next modeling attempts, this is an area for potential improvement.
- Linearity: The predicted vs. actual model shows a very good fit, and strong linear response characteristics. In addition, the residuals plot do not show any obvious tendency towards non-linearity or increasing / decreasing variance across the range of values evaluated. The assumption of linearity appears respected for this model.
- Equal variance: The residuals plot and the studentized residuals plots both show respect within reasonable bounds that equal variances here are acceptable.
- Cook's D and the Leverage Plots show a few influential points. The leverage plot, in particular, shows 2 points with relatively high leverage; however in this model, with 1400+ additional data points, the influence of these 2 points is not substantial.

Comparing Competing Models

Model	Adj R^2	CV Press	Kaggle Score
Forward	0.91	22.10	0.189
Backward	0.92	23.17	0.226
Stepwise	0.91	22.13	0.133
Custom	0.90		0.225

Conclusion:

- For this effort, the 4 models each provide good predictive capability for estimating market valuation of residential real estate to be proposed for offering in the Ames, Iowa market.
- The stepwise model outperformed the other selection methods for the features chosen in this case. In addition, the stepwise selection also resulted in the least number of features (14) in comparison, also providing a simpler model. The table above shows the characteristics relative to model fit, along with the Kaggle scores when the model is applied to the test case data set. Clearly, the stepwise model selection is the preferred model among those evaluated.

The retained features in the final model include :

Feature	Feature	Feature	Feature	Feature
bsmtfinsf1	centralair	fireplaces	garagecars	kitchenqual
log(grlivarea)	log(lotarea)	mszoning	neighborhood	overallcond
overallqual	totalbsmtsf	yearbuilt	yearremodadd	

Appendix

The code to complete this analysis can be found at this github site : https://github.com/bici-sancta/home_prices (https://github.com/bici-sancta/home_prices)

The complete set of plots used to downselect to the modeled features along with the residual plots for the remaining models are also found here : https://github.com/bici-sancta/home_prices/blob/master/homes_train_plots.pdf (https://github.com/bici-sancta/home_prices/blob/master/homes_train_plots.pdf)

The screen shot of the Kaggle scores associated to this model is found at this location : https://github.com/bici-sancta/home_prices/blob/master/kaggle_scores.png (https://github.com/bici-sancta/home_prices/blob/master/kaggle_scores.png)

SAS Codes

Question 1

```
SAS Code
FILENAME train '/home/pswaminathan0/ImportFiles/train.csv';

PROC IMPORT DATAFILE=train DBMS=CSV OUT=WORK.train replace;
    GETNAMES=YES;
RUN;

/*Filter data to include only specific neighboorhood and required columns*/
data train2;
    set train;
    keep ID Neighborhood SalePrice GrLivArea;
    If Neighborhood='BrkSide' OR Neighborhood='Edwards' or Neighborhood='NAme
s';
run;

/*Analysis of data*/
ODS GRAPHICS / ATTRPRIORITY=NONE;

/*to use different markers for groups rather than the default, which is diffe
rent colors; */
PROC SGPLOT DATA=train2;
    STYLEATTRS DATASYMBOLS=(Circle Triangle Asterisk);
    *sets symbols for groups (alphabetical);
    SCATTER X=GrLivArea Y=SalePrice/ GROUP=Neighborhood;
RUN;

proc sgscatter data=train2;
    matrix GrLivArea SalePrice;
run;

/*create dataset which log transforms sales price and living area
different combination of log transform is used to find the best model*/
data train3;
    set train2;
    logGrLivArea=log(GrLivArea);
    logSalePrice=log(SalePrice);
    GrLivArea_ft=GrLivArea/100;
    logGrLivArea_ft=log(GrLivArea/100);
run;

/**making brkside as ref has reduced the pvalue on edward;*/
proc glm data=train3 plots=all;
    class Neighborhood (REF='BrkSide');
    model logSalePrice=Neighborhood | logGrLivArea_ft / cli solution;
    output out=results p=Predict;
    run;
```

Question 2

(stewise selection, as that produced the best model of those considered)

```
proc datasets lib=work kill nolist memtype=data;
quit;

# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
=-=-
# ...       The cleaned data set was obtained by some R code processing
# ...
# ...       the R code for the data cleaning can be found at :
# ...       https://github.com/bici-sancta/home_prices/blob/master/home_price
s_data_prep.Rmd
# ...
# ...       the cleaned data set itself can be found at this location :
# ...       https://github.com/bici-sancta/home_prices/blob/master/data/train
ing_set_cleaned.csv
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
=-=-


FILENAME REFFILE '/folders/myfolders/training_set_cleaned.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT = training_set;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA = training_set; RUN;

/*
proc print data = training_set;
run;
*/

ods graphics on;

proc glmselect data = training_set plots = all;
class bsmtfintype1
    bsmtqual
    centralair
    electrical
    exterior1st
    exterior2nd
    exterqual
    fireplacequ
    foundation
    garagefinish
```

```
    garagetype
    heatingqc
    housestyle
    kitchenqual
    lotshape
    masvnrtype
    mszoning
    neighborhood
    saletype;
model log_saleprice =
            bsmtfinsf1
            bsmtfintype1
            bsmtfullbath
            bsmtqual
            centralair
            electrical
            exterior1st
            exterior2nd
            exterqual
            fireplacequ
            fireplaces
            foundation
            fullbath
            garagearea
            garagecars
            garagefinish
            garagetype
            halfbath
            heatingqc
            housestyle
            kitchenqual
            log_grlivarea
            log_lotarea
            log_lotfrontage
            lotshape
            masvnrtype
            mszoning
            neighborhood
            overallcond
            overallqual
            saletype
            totalbsmtsf
            totrmsabvgrd
            x1stflrsf
            x2ndflrsf
            yearbuilt
```

```
            yearremodadd
            /selection = stepwise(stop = cv) cvmethod = random(5) showpvalues
;
run;

ods graphics off;

/* features retained by stepwsie selection (2017.04.23) */

ods graphics on;
proc glm data = training_set plots=diagnostics;
class centralair
    kitchenqual
    mszoning
    neighborhood;
model log_saleprice =
    bsmtfinsf1
    centralair
    fireplaces
    garagecars
    kitchenqual
    log_grlivarea
    log_lotarea
    mszoning
    neighborhood
    overallcond
    overallqual
    totalbsmtsf
    yearbuilt
    yearremodadd
    /solution;
run;
ods graphics off;
```