BSAN 450 Take Home Exam

The data in the file named AmesHousing.csv is a data set that contains information from the Ames Assessor’s Office about individual residential properties sold in Ames, Iowa from 2006 to 2010. The purpose of this project is to develop a model for the Sale Price of a home in Ames based on the other variables in the data set using the statistical methods we have covered up to this point in the class. You are to complete this project on your own without anyone’s help. If you have questions about the project you can ask Professor Hillmer but you cannot ask anyone else.

The description of the variables is given at the end of this document.

For this project you need to do the following:

1) Develop a model to predict the variable SalePrice based on the other variables in the data set.

a) For each step of your development, clearly describe what you are doing and your reasons for taking this step. If you do not describe your reasoning you will lose points because I cannot read your mind. I need to know the rationale for what you are doing in developing the model.

b) When appropriate provide output from R to justify the conclusions you are making. I do not need to see everything you do in R, but I need enough output to follow your reasoning.

2) After you have developed a model, predict the SalePrice for the 2 houses that are in the file AmesHousing\_predict.csv. Compute a prediction along with 95% prediction intervals for the SalePrice.

Considerations in Grading:

Initial Model Determination: There are many possible independent variables in the data set. You should have a strategy for determining which of these variables you will include in the model and you should clearly describe that strategy so that I can understand your approach. You should provide an appropriate output from R to support the choices you make.

Model Checking and Modification: Once you have estimated a model, describe what you do to verify that the model assumptions are approximately correct and what changes to the model you are making. You should clearly describe why you are making changes in a model. Provide appropriate R output to support your choices.

Dealing with Outliers: In a large data set there are undoubtedly outliers. Describe how you identify any outliers and how you propose to deal with the observations that are outliers.

Model Verification: You should have a method to validate your model on an independent set of data. Describe how you propose to do this and provide an appropriate R output.

There are many possible models that can be derived from this data set, therefore there are many possible correct answers. I am not just interested in the final model, but also in the process your go through to arrive at your model. A significant amount of the grade for the project will be determined by the process you use to come up with your final model. If you have a poor process or do not clearly describe that process your grade will be reduced significantly.

Suggestions on how to approach the analysis of the data:

1) Since you will be doing a lot of things that involve using the data to suggest which variables to include in a model, it is a good idea to divide the data set into a training set and a test set. You should then do all of your analysis to determine a model on the training set without using the test set. After you have come up with a model using the training set you can then use the test set to check to see if all the variables that you have included in your model are needed.

2) The data set has a number of missing values. R has a variety of ways to deal with missing values and sometimes this can cause problems, thus when you are developing a model it is a good idea to remove the observations who have missing values. This seems OK with this data set because there are a large number of observations in the data set so that removing those with missing values will still leave a lot of data. The command to remove the missing values is: newdataframe = na.omit(dataframe) where newdataframe is the name you give to the data frame without the missing values and dataframe is the name of the data frame that has the data with missing values.

After you develop a model, it is a good idea to refit the model using the entire data set (including missing values) because your final model will probably not include some of the inputs which have missing values.

3) This is a data set with a very large number of possible input variables and in addition there are a large number of categorical variables. Because there are a large number of possible input variables, it is natural to try to use some sort of variable selection method like best subsets regression or stepwise regression to help in determining which input variables to choose. It can be a problem using a variable selection method with categorical variables because for a categorical variable with K categories, R will create K-1 indicator variables. This will greatly increase the number of potential input variables and could cause a problem.

One way to proceed is to break the data into 2 groups: the group of numeric variables and the group of categorical (factor) variables and to use a sequential approach to determining which variables to include in the model like that described below.

Step 1: Create a data frame that contains only the numeric inputs and the dependent variable, SalePrice. If you have read the data into a dataframe named ames, then a command to select these variables is:

num.ames=data.frame(ames[,c(2,3,14:17,23,31,33:35,40:49,51,53,56,58,59,63:68,70,71,74)])

This will create a dataframe named num.ames that only contains the variables in the columns listed in the command. These are the columns that contain the numeric variables and the variable SalePrice.

Step 2: Use the new data frame to develop a model that only includes the numeric variables. When you are satisfied that you have a good model for the numeric variables you can add the appropriate categorical variables. (Note, you should have removed the missing values.)

Step 3: You need a way to decide which categorical variables to add to the model you developed in Step 2. One way to do this is to start with the model you have from step 2, call this model 1, and then add one or more categorical variables to this model. Call the model with the added categorical variables model 2. Then the following R commands will perform a hypothesis test of whether or not the variables that are added to model 1 are statistically significant.

model1 = lm( the R expressions for model 1)

model2 = lm(the R expressions for model 2)

anova(model1,model2)

The output from the anova command will give a p-value for testing the null hypothesis that model1 is correct versus the alternative hypothesis that model2 is correct. The smaller the p-value, the more evidence in favor of model2.

You want to find a way to figure out which of the categorical variables are most likely to improve the model you developed in step 2. Once you have decided on which categorical variables are most likely to be helpful, you can add those to the model you developed in step 2 to get a model that includes both numeric and categorical variables.

4) When the data set has missing values, when R computes the residuals for a model, it does not compute the residuals for any data that is missing. Because of that the number of residuals is not the same as the number of variables in the data set. This causes a problem when you are trying to identify the outliers is a set of data that contains missing values. One way to resolve this problem is to include the expression na.action=na.exclude in the regression command. This will tell the program to put missing values in the residuals in places where the original data had missing data. The command would be like the following:

fit=lm(Y~X1+X2 + X3, data=dataframename, na.action=na.exclude)

If you use this command, then the residuals with large values will correspond to the rows of the observations that are associated with those large observations and you will be able to properly identify the observations corresponding to the outliers.

SOURCES:

Ames, Iowa Assessor’s Office

VARIABLE DESCRIPTIONS:

MS Zoning (Nominal): Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

Lot Frontage (Continuous): Linear feet of street connected to property

Lot Area (Continuous): Lot size in square feet

Street (Nominal): Type of road access to property

Grvl Gravel

Pave Paved

Alley (Nominal): Type of alley access to property

Grvl Gravel

Pave Paved

None No alley access

Lot Shape (Ordinal): General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

Land Contour (Nominal): Flatness of the property

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

Lot Config (Nominal): Lot configuration

Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property

Land Slope (Ordinal): Slope of property

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

Condition 1 (Nominal): Proximity to various conditions

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

Condition 2 (Nominal): Proximity to various conditions (if more than one is present)

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

Bldg Type (Nominal): Type of dwelling

1Fam Single-family Detached

2FmCon Two-family Conversion; originally built as one-family dwelling

Duplx Duplex

TwnhsE Townhouse End Unit

TwnhsI Townhouse Inside Unit

House Style (Nominal): Style of dwelling

1Story One story

1.5Fin One and one-half story: 2nd level finished

1.5Unf One and one-half story: 2nd level unfinished

2Story Two story

2.5Fin Two and one-half story: 2nd level finished

2.5Unf Two and one-half story: 2nd level unfinished

SFoyer Split Foyer

SLvl Split Level

Overall Qual (Ordinal): Rates the overall material and finish of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

Overall Cond (Ordinal): Rates the overall condition of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

Year Built (Discrete): Original construction date

Year Remod/Add (Discrete): Remodel date (same as construction date if no remodeling or additions)

Roof Style (Nominal): Type of roof

Flat Flat

Gable Gable

Gambrel Gabrel (Barn)

Hip Hip

Mansard Mansard

Shed Shed

Roof Matl (Nominal): Roof material

CompShg Standard (Composite) Shingle

Other Clay, Tile, Metal, Roll, Membrane, or Gravel & Tar

Wood Wood Shakes or Wood Shingles

Exterior 1 (Nominal): Exterior covering on house

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

Exterior 2 (Nominal): Exterior covering on house (if more than one material)

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

Mas Vnr Type (Nominal): Masonry veneer type

BrkCmn Brick Common

BrkFace Brick Face

None None

Stone Stone

Mas Vnr Area (Continuous): Masonry veneer area in square feet

Exter Qual (Ordinal): Evaluates the quality of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

Exter Cond (Ordinal): Evaluates the present condition of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

Foundation (Nominal): Type of foundation

BrkTil Brick & Tile

CBlock Cinder Block

PConc Poured Contrete

Slab Slab

Stone Stone

Wood Wood

Bsmt Qual (Ordinal): Evaluates the height of the basement

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (<79 inches)

None No Basement

Bsmt Cond (Ordinal): Evaluates the general condition of the basement

Gd Good or Excellent

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling or Poor - Severe cracking, settling, or wetness

None No Basement

Bsmt Exposure (Ordinal): Refers to walkout or garden level walls

Gd Good Exposure

Av Average Exposure (split levels or foyers typically score average or above)

Mn Mimimum Exposure

No No Exposure

None No Basement

BsmtFin Type 1 (Ordinal): Rating of basement finished area

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

None No Basement

BsmtFin SF 1 (Continuous): Type 1 finished square feet

BsmtFinType 2 (Ordinal): Rating of basement finished area (if multiple types)

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

None No Basement

BsmtFin SF 2 (Continuous): Type 2 finished square feet

Bsmt Unf SF (Continuous): Unfinished square feet of basement area

Total Bsmt SF (Continuous): Total square feet of basement area

Heating (Nominal): Type of heating

Gas Gas forced warm air furnace or Gas hot water or steam heat

Other Hot water or steam heat other than gas or Wall furnace or Gravity furnace or Floor Furnace

HeatingQC (Ordinal): Heating quality and condition

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

Central Air (Nominal): Central air conditioning

N No

Y Yes

Electrical (Ordinal): Electrical system

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

1st Flr SF (Continuous): First Floor square feet

2nd Flr SF (Continuous) : Second floor square feet

Low Qual Fin SF (Continuous): Low quality finished square feet (all floors)

Gr Liv Area (Continuous): Above grade (ground) living area square feet

Bsmt Full Bath (Discrete): Basement full bathrooms

Bsmt Half Bath (Discrete): Basement half bathrooms

Full Bath (Discrete): Full bathrooms above grade

Half Bath (Discrete): Half baths above grade

Bedroom (Discrete): Bedrooms above grade (does NOT include basement bedrooms)

Kitchen (Discrete): Kitchens above grade

KitchenQual (Ordinal): Kitchen quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

TotRmsAbvGrd (Discrete): Total rooms above grade (does not include bathrooms)

Functional (Ordinal): Home functionality (Assume typical unless deductions are warranted)

Typ Typical Functionality

Min1 Minor Deductions 1

Min2 Minor Deductions 2

Mod Moderate Deductions

Maj1 Major Deductions 1

Maj2 Major Deductions 2

Sev Severely Damaged

Sal Salvage only

Fireplaces (Discrete): Number of fireplaces

FireplaceQu (Ordinal): Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or

Garage Type (Nominal): Garage location

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

BuiltIn Built-In (Garage part of house - typically has room above garage)

CarPort Car Port

Detchd Detached from home

None No Garage

Garage Yr Blt (Discrete): Year garage was built

Garage Finish (Ordinal) : Interior finish of the garage

Fin Finished

RFn Rough Finished

Unf Unfinished

None No Garage

Garage Cars (Discrete): Size of garage in car capacity

Garage Area (Continuous): Size of garage in square feet

Garage Qual (Ordinal): Garage quality

Gd Good or Excellent

TA Typical/Average

Fa Fair or Poor

None No Garage

Garage Cond (Ordinal): Garage condition

Gd Good or Excellent

TA Typical/Average

Fa Fair or Poor

None No Garage

Paved Drive (Ordinal): Paved driveway

Y Paved

P Partial Pavement

N Dirt/Gravel

Wood Deck SF (Continuous): Wood deck area in square feet

Open Porch SF (Continuous): Open porch area in square feet

Enclosed Porch (Continuous): Enclosed porch area in square feet

3-Ssn Porch (Continuous): Three season porch area in square feet

Screen Porch (Continuous): Screen porch area in square feet

Pool Area (Continuous): Pool area in square feet

Fence (Ordinal): Fence quality

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

None No Fence

Misc Val (Continuous): $Value of miscellaneous feature

Yr Sold (Discrete): Year Sold (YYYY)

Sale Type (Nominal): Type of sale

WD Warranty Deed - Conventional

CWD Warranty Deed - Cash

VWD Warranty Deed - VA Loan

New Home just constructed and sold

COD Court Officer Deed/Estate

Con Contract 15% Down payment regular terms

ConLw Contract Low Down payment and low interest

ConLI Contract Low Interest

ConLD Contract Low Down

Oth Other

Sale Condition (Nominal): Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit

Family Sale between family members

Partial Home was not completed when last assessed (associated with New Homes)

SalePrice (Continuous): Sale price $$