Housing Prices

# Introduction

Many factors affect housing prices. In fact, there are so many variables that affect housing prices, it can be difficult to know if a specific variable is important for a specific house in a specific neighborhood. However it is possible to construct a statistical model to help find which variables are important and to what degree.

# Data Description

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

\* train has 1460 obs and 81 columns

\* test has 1459 obs 80 columns

# Exploratory Data Analysis

# Appendix

/\*

\* Import the datasets

\* train has 1460 obs and 81 columns

\* test has 1459 obs 80 columns

Char vars:

Alley BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtQual CentralAir Condition1 Condition2 Electrical ExterCond ExterQual Exterior1st Exterior2nd Fence FireplaceQu Foundation Functional GarageCond GarageFinish GarageQual GarageType Heating HeatingQC HouseStyle KitchenQual LandContour LandSlope LotConfig LotShape MSZoning MasVnrType MiscFeature Neighborhood PavedDrive PoolQC RoofMatl RoofStyle SaleCondition SaleType Street Utilities

Num vars:

BedroomAbvGr BsmtFinSF1 BsmtFinSF2 BsmtFullBath BsmtHalfBath BsmtUnfSF EnclosedPorch Fireplaces FullBath GarageArea GarageCars GrLivArea HalfBath KitchenAbvGr LotArea LowQualFinSF MSSubClass MiscVal MoSold OpenPorchSF OverallCond OverallQual PoolArea ScreenPorch TotRmsAbvGrd TotalBsmtSF WoodDeckSF YearBuilt YearRemodAdd YrSold \_1stFlrSF \_2ndFlrSF \_3SsnPorch

Extra vars:

Id

response variable:

SalePrice

\*/

**proc** **import**

datafile='/home/iangelov0/kaggle/train.csv'

out=train

dbms=CSV

replace;

getnames=yes;

datarow=**2**;

guessingrows=**2000**;

**proc** **import**

datafile='/home/iangelov0/kaggle/test.csv'

out=test

dbms=CSV

replace;

getnames=yes;

datarow=**2**;

guessingrows=**2000**;

/\*

\* Some vars in the test dataset have char data type instead of numeric

\* This will change their data type

\* Ugly but I don't know a better way to change data types

\*/

**data** test;

set test;

new1 = input(BsmtFinSF1, **8.**);

drop BsmtFinSF1;

if new1=. then new1=**0**;

rename new1=BsmtFinSF1;

new2 = input(BsmtFinSF2, **8.**);

drop BsmtFinSF2;

if new2=. then new2=**0**;

rename new2=BsmtFinSF2;

new3 = input(BsmtUnfSF, **8.**);

drop BsmtUnfSF;

if new3=. then new3=**0**;

rename new3=BsmtUnfSF;

new4 = input(TotalBsmtSF, **8.**);

drop TotalBsmtSF;

if new4=. then new4=**0**;

rename new4=TotalBsmtSF;

new5 = input(BsmtFullBath, **8.**);

drop BsmtFullBath;

if new5=. then new5=**0**;

rename new5=BsmtFullBath;

new6 = input(BsmtHalfBath, **8.**);

drop BsmtHalfBath;

if new6=. then new6=**0**;

rename new6=BsmtHalfBath;

if new6=. then new6=**0**;

rename new6=BsmtHalfBath;

new7 = input(GarageArea, **8.**);

drop GarageArea;

if new7=. then new7=**0**;

rename new7=GarageArea;

new8 = input(GarageCars, **8.**);

drop GarageCars;

if new8=. then new8=**0**;

rename new8=GarageCars;

/\*

\* Append the test dataset on the end of train

\*/

**data** dataset;

set train test;

/\*

\* Drop problematic vars

\*/

**data** dataset;

set dataset;

drop GarageYrBlt LotFrontage MasVnrArea;

/\*

\* Replace character missing values with 'NA'

\*/

**data** dataset;

set dataset;

array change \_character\_;

do over change;

if missing(change) then change='NA';

end;

if missing(change) then change='NA';

end;

/\*

\* Output some data

\*/

**proc** **print** data=dataset (obs=**10**);

**proc** **means** data=dataset;

**proc** **means** data=dataset N NMISS;

/\*

\* Variable transformations of the continious vars

\*/

/\* LotAreaTrans OverallQualTrans OverallCondTrans YearRemodAddTrans BsmtFinSF1 BsmtFinSF1Flag \*/

/\* TotalBsmtSFTrans TotalBsmtSFFlag \_1stFlrSFTrans \_2ndFlrSFFlag \_2ndFlrSFFlag GrLivAreaTrans \*/

/\* BsmtFullBath FullBath HalfBath BedroomAbvGrTrans KitchenAbvGr TotRmsAbvGrd Fireplaces \*/

/\* GarageCars GarageArea WoodDeckSF WoodDeckSFFlag \*/

**data** dataset;

set dataset;

SalePriceSqrt=sqrt(SalePrice);

drop SalePrice;

LotAreaTrans = log(LotArea);

OverallQualTrans = log(OverallQual);

OverallCondTrans = log(OverallCond);

YearRemodAddTrans = log(YearRemodAdd );

BsmtFinSF1Flag = **0**;

if BsmtFinSF1 = **0** then do;

BsmtFinSF1 = **700**;

BsmtFinSF1Flag = **1**;

end;

TotalBsmtSFTrans = sqrt(TotalBsmtSF);

TotalBsmtSFFlag = **0**;

if TotalBsmtSFTrans = **0** then do;

TotalBsmtSFTrans = sqrt(TotalBsmtSF);

TotalBsmtSFTrans = **30**;

TotalBsmtSFFlag = **1**;

end;

\_1stFlrSFTrans = log(\_1stFlrSF);

\_2ndFlrSFFlag = **0**;

if \_2ndFlrSF = **0** then do;

\_2ndFlrSF = **750**;

\_2ndFlrSFFlag = **1**;

end;

GrLivAreaTrans = log(GrLivArea);

BedroomAbvGrTrans = log(BedroomAbvGr);

WoodDeckSFFlag = **0**;

if WoodDeckSF = **0** then do;

WoodDeckSF = **200**;

WoodDeckSFFlag = **1**;

end;

keep Id SalePriceSqrt LotAreaTrans OverallQualTrans OverallCondTrans YearRemodAddTrans BsmtFinSF1 BsmtFinSF1Flag TotalBsmtSFTrans TotalBsmtSFFlag \_1stFlrSFTrans \_2ndFlrSFFlag \_2ndFlrSFFlag GrLivAreaTrans BsmtFullBath FullBath HalfBath BedroomAbvGrTrans KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF WoodDeckSFFlag Alley BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtQual CentralAir Condition1 Condition2 Electrical ExterCond ExterQual Exterior1st Exterior2nd Fence FireplaceQu Foundation Functional GarageCond GarageFinish GarageQual GarageType Heating HeatingQC HouseStyle KitchenQual LandContour LandSlope LotConfig LotShape MSZoning MasVnrType MiscFeature Neighborhood PavedDrive PoolQC RoofMatl RoofStyle SaleCondition SaleType Street Utilities;

/\*

\* Create dummy variables

\*/

**proc** **glmmod** data=dataset outdesign=dummies noprint;

class Alley BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtQual CentralAir Condition1 Condition2 Electrical ExterCond ExterQual Exterior1st Exterior2nd Fence FireplaceQu Foundation Functional GarageCond GarageFinish GarageQual GarageType Heating HeatingQC HouseStyle KitchenQual LandContour LandSlope LotConfig LotShape MSZoning MasVnrType MiscFeature Neighborhood PavedDrive PoolQC RoofMatl RoofStyle SaleCondition SaleType Street Utilities;

model Id = Alley BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtQual CentralAir Condition1 Condition2 Electrical ExterCond ExterQual Exterior1st Exterior2nd Fence FireplaceQu Foundation Functional GarageCond GarageFinish GarageQual GarageType Heating HeatingQC HouseStyle KitchenQual LandContour LandSlope LotConfig LotShape MSZoning MasVnrType MiscFeature Neighborhood PavedDrive PoolQC RoofMatl RoofStyle SaleCondition SaleType Street Utilities;

/\*

\* Merge the dummy vars with the dataste

\*/

**data** dataset;

merge dataset dummies;

/\*

\* Don't need the char vars anymore because they are represented by dummies

\* This will only keep numerical variables in the dataset

\*/

**data** dataset;

set dataset;

KEEP \_numeric\_;

**proc** **contents** data=dataset;

/\*

\* Get cooks'd for all continious vars

\* influance will contain the result

\*/

**proc** **reg** data=dataset plots=diagnostics;

model SalePriceSqrt=BsmtFinSF1 BsmtFullBath FullBath HalfBath KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF LotAreaTrans OverallQualTrans OverallCondTrans YearRemodAddTrans BsmtFinSF1Flag TotalBsmtSFTrans TotalBsmtSFFlag \_1stFlrSFTrans \_2ndFlrSFFlag GrLivAreaTrans BedroomAbvGrTrans WoodDeckSFFlag

/ noprint;

output out=influance cookd=cookd NOPRINT;

/\*

\* Remove all columns but cookd

\* Just to keep the dataset clean

\*/

**data** influance;

set influance;

keep cookd;

/\*

\* Merge the two datasets

\* dataset will now have a variable cookd

\*/

**data** dataset;

\*/

**data** dataset;

merge dataset influance;

/\*

\* Remove all very large cooks'd

\* A rule of thumb cutoff is 4/N (N - obs)

\*/

**data** dataset;

set dataset;

where cookd < **4**/**1460**;

drop cookd;

**proc** **print** data=dataset (obs=**10**);

**proc** **means** data=dataset;

/\*

\* Get only train part of the data

\*/

**data** train;

set dataset;

where SalePriceSqrt ^= .;

/\*

\* Variable Selection

\* other methods: LASSO stepwise(choose=BIC) selection=LASSO(stop=CV) cvMethod=RANDOM(20)\*

\*/

**proc** **glmselect** data=train;

model SalePriceSqrt=

BsmtFinSF1 BsmtFullBath FullBath HalfBath KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF LotAreaTrans OverallQualTrans OverallCondTrans YearRemodAddTrans BsmtFinSF1Flag TotalBsmtSFTrans TotalBsmtSFFlag \_1stFlrSFTrans \_2ndFlrSFFlag GrLivAreaTrans BedroomAbvGrTrans WoodDeckSFFlag Col1 Col2 Col3 Col4 Col5 Col6 Col7 Col8 Col9 Col10 Col11 Col12 Col13 Col14 Col15 Col16 Col17 Col18 Col19 Col20 Col21 Col22 Col23 Col24 Col25 Col26 Col27 Col28 Col29 Col30 Col31 Col32 Col33 Col34 Col35 Col36 Col37 Col38 Col39 Col40 Col41 Col42 Col43 Col44 Col45 Col46 Col47 Col48 Col49 Col50 Col51 Col52 Col53 Col54 Col55 Col56 Col57 Col58 Col59 Col60 Col61 Col62 Col63 Col64 Col65 Col66 Col67 Col68 Col69 Col70 Col71 Col72 Col73 Col74 Col75 Col76 Col77 Col78 Col79 Col80 Col81 Col82 Col83 Col84 Col85 Col86 Col87 Col88 Col89 Col90 Col91 Col92 Col93 Col94 Col95 Col96 Col97 Col98 Col99 Col100 Col101 Col102 Col103 Col104 Col105 Col106 Col107 Col108 Col109 Col110 Col111 Col112 Col113 Col114 Col115 Col116 Col117 Col118 Col119 Col120 Col121 Col122 Col123 Col124 Col125 Col126 Col127 Col128 Col129 Col130 Col131 Col132 Col133 Col134 Col135 Col136 Col137 Col138 Col139 Col140 Col141 Col142 Col143 Col144 Col145 Col146 Col147 Col148 Col149 Col150 Col151 Col152 Col153 Col154 Col155 Col156 Col157 Col158 Col159 Col160 Col161 Col162 Col163 Col164 Col165 Col166 Col167 Col168 Col169 Col170 Col171 Col172 Col173 Col174 Col175 Col176 Col177 Col178 Col179 Col180 Col181 Col182 Col183 Col184 Col185 Col186 Col187 Col188 Col189 Col190 Col191 Col192 Col193 Col194 Col195 Col196 Col197 Col198 Col199 Col200 Col201 Col202 Col203 Col204 Col205 Col206 Col207 Col208 Col209 Col210 Col211 Col212 Col213 Col214 Col215 Col216 Col217 Col218 Col219 Col220 Col221 Col222 Col223 Col224 Col225 Col226 Col227 Col228 Col229 Col230 Col231 Col232 Col233 Col234 Col235 Col236 Col237 Col238 Col239 Col240 Col241 Col242 Col243 Col244 Col245 Col246 Col247 Col248 Col249 Col250 Col251 Col252 Col253 Col254 Col255 Col256 Col257 Col258 Col259 Col260 Col261 Col262 Col263 Col264 Col265 Col266 Col267 Col268 Col269 Col270 Col271 Col272 Col273 Col274 Col275 Col276

/ selection=LASSO(stop=CV) cvMethod=RANDOM(**3**);

/\*

\* Fit the linear regression

\* and predict SalePriceSqrt for the test part of the dataset

\*/

**proc** **glm** data=dataset;

model SalePriceSqrt=BsmtFinSF1 BsmtFullBath FullBath HalfBath KitchenAbvGr Fireplaces GarageCars GarageArea WoodDeckSF LotAreaTrans OverallQualTrans OverallCondTrans YearRemodAddTrans BsmtFinSF1Flag TotalBsmtSFTrans GrLivAreaTrans WoodDeckSFFlag Col16 Col19 Col22 Col34 Col38 Col39 Col41 Col43 Col69 Col72 Col76 Col119 Col130 Col137 Col140 Col150 Col160 Col165 Col173 Col177 Col199 Col216 Col217 Col225 Col226 Col227 Col231 Col232 Col237 Col256 Col261;

output out=regout(where=(SalePriceSqrt=.)) p=predicted;

**proc** **print** data=regout (obs=**10**);

/\*

\* Create the final dataset

\*/

**data** submission;

set regout;

/\* since we took square root now have to addjuct back \*/

SalePrice = predicted \* predicted;

keep Id SalePrice;

**proc** **print** data=submission (obs=**10**);

/\*

\* Export the result to csv

\*/

**proc** **export** data=submission dbms=csv

outfile="/home/iangelov0/kaggle/submission.csv"

replace;

**run**;