01/08/2023

The (Numerical) Variables that correspond with one another closely:   
Training Data Notes:

* Gr. Living Area (GrLivArea: Above grade (ground) living area square feet) AND Total Rooms above Ground highly correlated at 0.81
* Year built and Garage Year Built obviously correspond highly at 0.83

Going for the jugular, in terms of `sales price` we have

Overall Quality 0.8

Gr. Living Early 0.7

Total Basement Square Footage at 0.63

1st floor sq ft at 0.62

Garage cars = garage area at 0.65

Test Data Notes:

The variables we’ll use (All the ones that are normally distributed as determined by histograms)  
  
Lot Frontage (lawn)

Overall Quality (Big R^2 Value)

Overall Condition

Year Built (Norm distributed but a ton of volume after 2000) [note a bunch of homes remodeled in the 2000s]

Gr. Living Area and Total Rooms above ground (correlate very strongly at 0.81)

Garage Area

These are the highest R^2:   
  
1st Flr SF 0.618486

Total Bsmt SF 0.628925

Garage Cars 0.648220

Garage Area 0.650270

Gr Liv Area 0.697038

Overall Qual 0.800207

Regarding Cleaning:

I’m using these:   
  
features = ['Overall Qual', 'Gr Liv Area', 'Garage Area', 'Garage Cars','Total Bsmt SF','1st Flr SF']  
  
^ Drop Nan’s from these

Linear is 0.7532949222487371

R^2 is up to 0.78 train

0.80 test across the models

01/09/2023

Train is 0.80

Test is 0.85

MAE

21716.11689230271

RMSE

27950.20013639609

Took out : 'Wood Deck SF','Overall Cond'

1/10:   
  
Sonyah:  
  
Explore more dummies variables

Activation Terms in Polynomial Features to raise Power (Keep Track of your Units)

Steer away from highest score

Look into coefficients and try to interpret coefficients

1/11

Check into Hot Encoders

Documenting what we’ve done so far:   
  
Polynomial features degree 2 is working the best, each consecutive increase in power of 1 saw drop off R^2 of 0.2 and 0.3 respectively.

Currently :   
  
Train: 0.92258

Test : 0.877

MAE

17440.858349246362

RMSE

25502.721480297838

Model is currently overly complex, so trying to remove some features, start with lower R^2 values, such as “Month Sold”  
  
'Mo Sold'

Removing brought them closer together but lowered R^2

The train score for ridge model is 0.9206022927870164

The test score for ridge model is 0.8833910498664316

MAE | 17256.74836531123

RMSE | 24930.295686239828

Good this is a step in the right direction

Try remove :

'Open Porch SF'

The train score for ridge model is 0.9180848374241882

The test score for ridge model is 0.8872874021934984

3.0797435230689785% difference

MAE | 17062.54936221826

RMSE | 24510.24862436753

Remove

'Gr Liv Area Times Garage Area'

The train score for ridge model is 0.9128627076851468

The test score for ridge model is 0.8838779348288656

2.8984772856281205% difference

MAE | 17341.270334806384

RMSE | 24878.194698597697

SO Put it back

Put back we have different alpha 475

The train score for ridge model is 0.9041834674849308

The test score for ridge model is 0.8847633257456613

1.9420141739269536% difference

MAE | 17324.6578691974

RMSE | 24783.16936746563 OG Columns:

['Id',

'PID',

'MS SubClass',

'MS Zoning',

'Lot Frontage',

'Lot Area',

'Street',

'Alley',

'Lot Shape',

'Land Contour',

'Utilities',

'Lot Config',

'Land Slope',

'Neighborhood',

'Condition 1',

'Condition 2',

'Bldg Type',

'House Style',

'Overall Cond',

'Roof Style',

'Roof Matl',

'Exterior 1st',

'Exterior 2nd',

'Mas Vnr Type',

'Exter Cond',

'Foundation',

'Bsmt Qual',

'Bsmt Cond',

'Bsmt Exposure',

'BsmtFin Type 1',

'BsmtFin Type 2',

'BsmtFin SF 2',

'Bsmt Unf SF',

'Heating',

'Heating QC',

'Central Air',

'Electrical',

'2nd Flr SF',

'Low Qual Fin SF',

'Bsmt Full Bath',

'Bsmt Half Bath',

'Half Bath',

'Bedroom AbvGr',

'Kitchen AbvGr',

'Kitchen Qual',

'Functional',

'Fireplace Qu',

'Garage Type',

'Garage Finish',

'Garage Qual',

'Garage Cond',

'Paved Drive',

'Wood Deck SF',

'Open Porch SF',

'Enclosed Porch',

'3Ssn Porch',

'Screen Porch',

'Pool Area',

'Pool QC',

'Fence',

'Misc Feature',

'Misc Val',

'Mo Sold',

'Yr Sold',

'Sale Type']

Added :

df['Quality Factor'] = df['Overall Qual'] \* df['Exter Qual']

kaggle\_data['Quality Factor'] = kaggle\_data['Overall Qual'] \* kaggle\_data['Exter Qual']

And new scores look good:   
  
The train score for ridge model is 0.9063250069759797

The test score for ridge model is 0.8877784091309094

1.85% difference

The train score for lin model is 0.9252269787345884

The test score for lin model is 0.876335458300263

4.89% difference

MAE Ridge| 17161.32150289649

RMSE Ridge | 24456.80367860701

Let’s keep going after we add ample documentation

New Alpha:   
  
ridge = Ridge(alpha=599.4842503189409)  
The train score for ridge model is 0.903652260939732

The test score for ridge model is 0.8866015160085979

1.71% difference

The train score for lin model is 0.9252269787345884

The test score for lin model is 0.876335458300263

4.89% difference

MAE Ridge| 17227.407091945734

RMSE Ridge | 24584.71121128331

Check with PID 527452310 abbrev: 2376

Markdown utilized successfully. We need to further document our work. Particularly the polynomial characteristic equation and as this increases holding constant business etc.

Yet to do:

More visualizations

Finish Presentation

ReadME start and finish

Completely Optimize to WIN

01/13/2023

Adding Dummies for kitchen Need all Dummy Variables for the attribute (Kitchen was first) to make error go back down (stabilize) and to make train and test converge. Here are metrics for adding :

'Kitchen Qual\_Ex','Kitchen Qual\_Fa','Kitchen Qual\_Gd','Kitchen Qual\_TA'

Ridge alpha: 599.4842503189409

The train score for ridge model is 0.9119171244570627

The test score for ridge model is 0.8838897899920699

2.8% difference

The train score for lin model is 0.9323566456712762

The test score for lin model is 0.8807273197938544

5.16% difference

MAE Ridge| 17221.837946038962

RMSE Ridge | 24876.924730715295

Next implementation wave is heating quality:   
  
'Heating QC\_Ex','Heating QC\_Fa','Heating QC\_Gd','Heating QC\_TA'

673.4150657750828

The train score for ridge model is 0.9135889205998982

The test score for ridge model is 0.8885029813351292

2.51% difference

The train score for lin model is 0.9362158154459609

The test score for lin model is 0.8856694866417067

5.05% difference

MAE Ridge| 16844.525222603308

RMSE Ridge | 24377.72167344444

Third wave was electrical fuses:

'Electrical\_FuseA','Electrical\_FuseF','Electrical\_FuseP','Electrical\_SBrkr'

We did worse:   
  
673.4150657750828

The train score for ridge model is 0.9146847999442931

The test score for ridge model is 0.8845278491906279

3.02% difference

The train score for lin model is 0.9380675220566039

The test score for lin model is 0.8631892865790636

7.49% difference

MAE Ridge| 17293.259271496485

RMSE Ridge | 24808.47761743947

I was misunderstanding the rule, so I filled NaN columns with mean values.

Now my scores are following:

KAGGLE RMSE: 25724.21379 TOP OF THE BOARD

That one was fill Lot and BSMT FL 1 with 0s

This one is fill with mean:

alpha

385.3528593710531

The train score for ridge model is 0.9192398897723265

The test score for ridge model is 0.9115037690974639

0.77% difference

The train score for lin model is 0.9348859550543305

The test score for lin model is 0.915022738138772

1.99% difference

MAE Ridge| 15554.663396247124

RMSE Ridge | 22825.08742640468

features = ['Overall Qual', 'Gr Liv Area', 'Garage Area', 'Garage Cars','Total Bsmt SF',

'1st Flr SF','Year Built','Year Remod/Add','Full Bath','Garage Yr Blt','TotRms AbvGrd',

'Mas Vnr Area','Fireplaces','BsmtFin SF 1','Wood Deck SF','Open Porch SF',

'After 2000','Exter Qual','Gr Liv Area Times Garage Area','Overall Qual',

# 'Kitchen Qual\_Ex','Kitchen Qual\_Fa','Kitchen Qual\_Gd','Kitchen Qual\_TA',

# 'Heating QC\_Ex','Heating QC\_Fa','Heating QC\_Gd','Heating QC\_TA'

]

X = df[features]

y = df['SalePrice']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, random\_state=42,

train\_size=0.90)

alpha

83.02175681319744

The train score for ridge model is 0.9218431653137447

The test score for ridge model is 0.9187480540107409

0.31% difference

The train score for lin model is 0.9278011011703752

The test score for lin model is 0.9176136289404316

1.02% difference

MAE Ridge| 15453.347186292094

RMSE Ridge | 21870.914767727398

Instead of filling with Mean for columns with Null; try to fill with Median

The train score for ridge model is 0.9299566037916331

The test score for ridge model is 0.920353157467107

0.96% difference

RMSE Ridge | 21653.811125072734

features = ['Overall Qual', 'Gr Liv Area', 'Garage Area', 'Garage Cars','Total Bsmt SF',

'1st Flr SF','Year Built','Year Remod/Add','Full Bath','Garage Yr Blt','TotRms AbvGrd',

'Mas Vnr Area','Fireplaces','BsmtFin SF 1','Wood Deck SF','Open Porch SF',

'After 2000','Exter Qual','Overall Qual','Gr Liv Area Times Garage Area',

'Kitchen Qual\_Ex','Kitchen Qual\_Fa','Kitchen Qual\_Gd','Kitchen Qual\_TA','Total Bsmt SF Times 1st Flr SF',

'Heating QC\_Ex','Heating QC\_Fa','Heating QC\_Gd','Heating QC\_TA',

# 'Neighborhood\_NridgHt','Neighborhood\_Timber','Bsmt Full Bath','Neighborhood\_StoneBr',

# 'Lot Frontage','Neighborhood\_Somerst','Neighborhood\_NoRidge'

]

**422.9242874389499**

**The train score for ridge model is 0.9263867855343225**

**The test score for ridge model is 0.9243491533316479**

**0.2% difference**

**RMSE Ridge | 21103.620038900877**

**Good Kaggle: 24618.45763 Score**

features = ['Overall Qual', 'Gr Liv Area', 'Garage Area', 'Garage Cars','Total Bsmt SF',

'1st Flr SF','Year Built','Year Remod/Add','Full Bath','Garage Yr Blt','TotRms AbvGrd',

'Mas Vnr Area','Fireplaces','BsmtFin SF 1','Wood Deck SF','Open Porch SF',

'After 2000','Exter Qual','Overall Qual','Gr Liv Area Times Garage Area',

'Kitchen Qual\_Ex','Kitchen Qual\_Fa','Kitchen Qual\_Gd','Kitchen Qual\_TA','Total Bsmt SF Times 1st Flr SF',

'Heating QC\_Ex','Heating QC\_Fa','Heating QC\_Gd','Heating QC\_TA','1st Flr SF Times 2nd Flr SF'

# 'Neighborhood\_NridgHt','Neighborhood\_Timber','Bsmt Full Bath','Neighborhood\_StoneBr',

# 'Lot Frontage','Neighborhood\_Somerst','Neighborhood\_NoRidge'

]

The train score for ridge model is 0.9271838171849146

The test score for ridge model is 0.9300720217443794

-0.29% difference

RMSE Ridge | 20289.69640040741

features = ['Overall Qual', 'Gr Liv Area', 'Garage Area', 'Garage Cars','Total Bsmt SF',

'1st Flr SF','Year Built','Year Remod/Add','Full Bath','Garage Yr Blt','TotRms AbvGrd',

'Mas Vnr Area','Fireplaces','BsmtFin SF 1','Wood Deck SF','Open Porch SF',

'After 2000','Exter Qual','Overall Qual','Gr Liv Area Times Garage Area',

'Kitchen Qual\_Ex','Kitchen Qual\_Fa','Kitchen Qual\_Gd','Kitchen Qual\_TA','Total Bsmt SF Times 1st Flr SF',

'Heating QC\_Ex','Heating QC\_Fa','Heating QC\_Gd','Heating QC\_TA','1st Flr SF Times 2nd Flr SF',

'Neighborhood\_NridgHt','Neighborhood\_Timber','Bsmt Full Bath','Neighborhood\_StoneBr',

'Neighborhood\_Somerst','Neighborhood\_NoRidge'

]

The train score for ridge model is 0.9356389910133571

The test score for ridge model is 0.9280341071727103

0.76% difference

RMSE Ridge | 20583.22499800934

# After 2000 is our only dummy variable

# Exter Qual is our only feature

#'Gr Liv Area Times Garage Area' is an area booster

# 'Total Bsmt SF Times 1st Flr SF' performed poorly and was deducted

#Trying Dummies 01/13/2023

#First Kitchen Quality

#Second Heating Quality

#Third Electrical Fuses - Do Not Use

#Now try the externals:

# take out 'Lot Frontage'

# Now try the inverses

features = ['Overall Qual', 'Gr Liv Area', 'Garage Area', 'Garage Cars','Total Bsmt SF',

'1st Flr SF','Year Built','Year Remod/Add','Full Bath','Garage Yr Blt','TotRms AbvGrd',

'Mas Vnr Area','Fireplaces','BsmtFin SF 1','Wood Deck SF','Open Porch SF',

'After 2000','Exter Qual','Overall Qual','Gr Liv Area Times Garage Area',

'Kitchen Qual\_Ex','Kitchen Qual\_Fa','Kitchen Qual\_Gd','Kitchen Qual\_TA','Total Bsmt SF Times 1st Flr SF',

'Heating QC\_Ex','Heating QC\_Fa','Heating QC\_Gd','Heating QC\_TA','1st Flr SF Times 2nd Flr SF',

'Neighborhood\_NridgHt','Neighborhood\_Timber','Bsmt Full Bath','Neighborhood\_StoneBr',

'Neighborhood\_Somerst','Neighborhood\_NoRidge',

# 'Functional\_Maj1','Functional\_Maj2','Functional\_Min1','Functional\_Min2','Functional\_Mod','Functional\_Typ'

'Not Kitchen Qual\_TA'

]

X = df[features]

y = df['SalePrice']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, random\_state=42,

train\_size=0.90)

The train score for ridge model is 0.9356277493246843

The test score for ridge model is 0.927864502510424

0.78% difference

RMSE Ridge | 20607.465345479737

features = ['Overall Qual', 'Gr Liv Area', 'Garage Area', 'Garage Cars','Total Bsmt SF',

'1st Flr SF','Year Built','Year Remod/Add','Full Bath','Garage Yr Blt','TotRms AbvGrd',

'Mas Vnr Area','Fireplaces','BsmtFin SF 1','Wood Deck SF','Open Porch SF',

'After 2000','Exter Qual','Overall Qual','Gr Liv Area Times Garage Area',

'Kitchen Qual\_Ex','Kitchen Qual\_Fa','Kitchen Qual\_Gd','Kitchen Qual\_TA','Total Bsmt SF Times 1st Flr SF',

'Heating QC\_Ex','Heating QC\_Fa','Heating QC\_Gd','Heating QC\_TA','1st Flr SF Times 2nd Flr SF',

'Neighborhood\_NridgHt','Neighborhood\_Timber','Bsmt Full Bath','Neighborhood\_StoneBr',

'Neighborhood\_Somerst','Neighborhood\_NoRidge',

# 'Functional\_Maj1','Functional\_Maj2','Functional\_Min1','Functional\_Min2','Functional\_Mod','Functional\_Typ'

'Not Kitchen Qual\_TA','Not Heating QC\_TA','Not Neighborhood\_OldTown'

]

The train score for ridge model is 0.9365907384851683

The test score for ridge model is 0.9288078538373481

0.78% difference

[2433]:

RMSE Ridge | 20472.274930670817

On Kaggle 23751.33991 . More tomorrow. And work on the ACTUAL project. More documentation more takeaways finish presentation

01/14/2023 more inverse work

Not much luck that day back to what we did before

Current parameters

features = ['Overall Qual', 'Gr Liv Area', 'Garage Area', 'Garage Cars','Total Bsmt SF',

'1st Flr SF','Year Built','Year Remod/Add','Full Bath','Garage Yr Blt','TotRms AbvGrd',

'Mas Vnr Area','Fireplaces','BsmtFin SF 1','Wood Deck SF','Open Porch SF',

'After 2000','Exter Qual','Overall Qual','Gr Liv Area Times Garage Area',

'Kitchen Qual\_Ex','Kitchen Qual\_Fa','Kitchen Qual\_Gd','Kitchen Qual\_TA','Total Bsmt SF Times 1st Flr SF',

'Heating QC\_Ex','Heating QC\_Fa','Heating QC\_Gd','Heating QC\_TA','1st Flr SF Times 2nd Flr SF',

'Neighborhood\_NridgHt','Neighborhood\_Timber','Bsmt Full Bath','Neighborhood\_StoneBr',

'Neighborhood\_Somerst','Neighborhood\_NoRidge',

# 'Functional\_Maj1','Functional\_Maj2','Functional\_Min1','Functional\_Min2','Functional\_Mod','Functional\_Typ'

'Not Kitchen Qual\_TA','Not Heating QC\_TA','Not Neighborhood\_OldTown'

]

Lasso just did well:   
  
**The train score for lasso model is 0.9282990792350672**

**The test score for lasso model is 0.927615791511294**

**0.07% difference**

**MAE Ridge | 15345.579003147821**

**MAE Lasso | 14448.11988434666**

**RMSE Ridge | 20622.708586182438**

**RMSE Lasso | 20642.96030481706**

**Kaggle score : 23322.66654 highest yet proceed with Lasso**  
  
01/15/2023

Took out External quality:

# After 2000 is our only dummy variable

# Exter Qual is our only feature

#'Gr Liv Area Times Garage Area' is an area booster

# 'Total Bsmt SF Times 1st Flr SF' performed poorly and was deducted

#Trying Dummies 01/13/2023

#First Kitchen Quality

#Second Heating Quality

#Third Electrical Fuses - Do Not Use

#Now try the externals:

# take out 'Lot Frontage'

# Now try the inverses

features = ['Overall Qual', 'Gr Liv Area', 'Garage Area', 'Garage Cars','Total Bsmt SF',

'1st Flr SF','Year Built','Year Remod/Add','Full Bath','Garage Yr Blt','TotRms AbvGrd',

'Mas Vnr Area','Fireplaces','BsmtFin SF 1','Wood Deck SF','Open Porch SF',

'After 2000','Overall Qual','Gr Liv Area Times Garage Area',

'Kitchen Qual\_Ex','Kitchen Qual\_Fa','Kitchen Qual\_Gd','Kitchen Qual\_TA','Total Bsmt SF Times 1st Flr SF',

'Heating QC\_Ex','Heating QC\_Fa','Heating QC\_Gd','Heating QC\_TA','1st Flr SF Times 2nd Flr SF',

'Neighborhood\_NridgHt','Neighborhood\_Timber','Bsmt Full Bath','Neighborhood\_StoneBr',

'Neighborhood\_Somerst','Neighborhood\_NoRidge',

'Functional\_Maj1','Functional\_Maj2','Functional\_Min1','Functional\_Min2','Functional\_Mod','Functional\_Typ',

'Not Kitchen Qual\_TA','Not Heating QC\_TA','Not Neighborhood\_OldTown','Not Heating QC\_Gd','Not Kitchen Qual\_Fa',

# 'Exterior 2nd\_VinylSd','Exterior 1st\_VinylSd','Exterior 2nd\_CmentBd','Exterior 1st\_CemntBd'

]

X = df[features]

y = df['SalePrice']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, random\_state=42,

train\_size=0.90)