Using regression models to predict housing prices

DSI 18

Robby, Mak, Ben D, Sahaj

Background and Context

Executive Summary

Context

- As employees in Iowa Real Estate Company, we are pitching our services to potential home-sellers in Ames City to consider our services to get the best value out of your home

Model & Conclusions

- Our housing dataset that contains over 2,400 transactions, we identified the core features that will better predict house prices for you
- Our model generates good predictive ability for house valuations up till $320,000 and subsequently tend to under-predict valuations for prices above this range

Workflow pipeline

1. Data cleaning: Removing outliers, standardising categorical variables.
2. Exploratory Data Analysis: Check correlations to guide initial hypothesis and feature selection.
3. Feature Engineering: Reducing the noise and amplifying the signal
4. Model Iteration and Selection: Preparing data with train/validate splits, then running ridge/lasso/elasticnet models for further feature selection. Models were compared on RMSE scores.
5. Model Evaluation: Understand what's working and what can be improved for the model, along with any caveats.

Getting to know Ames:

Feature	Values
Population	66,258
Area	71.2 sq km
Median Housing Value	$196,400
No. of Houses	26,754
No. Students in Iowa State Uni	33,391
Top Employer	16,811

Ames: The 9th best city to live in

BEST PLACES TO LIVE *Money's list of America's best small cities*

Beautiful houses in every neighborhood

North Ridge Heights

Stonebridge

Somerset

College Creek

Data Cleaning

Data Cleaning - Handling of Null Data

Drop those columns with more than 1000 Null values. i.e. `Alley`, `Fireplace Qu`, `Pool QC`, `Fence` & `Misc Feature`

In [11]:
```
# Initial data size
train.shape
```

Out[11]: (2049, 81)

In [12]:
```
# Dropping the five columns
train.drop(columns=['Alley' , 'Fireplace Qu' , 'Pool QC' , 'Fence' , 'Misc Feature'],axis=1,inplace=True)
```

In [13]:
```
# New data size
train.shape
```

Out[13]: (2049, 76)

Step 1: Drop those columns with more than 1,000 null values

Data Cleaning - Handling of Null Data

Handling of `Mas Vnr Type` null data.

In [17]:
```
# Check the number of elements inside 'Mas Vnr Type'
train['Mas Vnr Type'].value_counts(dropna=False)
```

Out[17]:
```
None       1218
BrkFace     630
Stone       166
NaN          22
BrkCmn       13
Name: Mas Vnr Type, dtype: int64
```

In [18]:
```
# Replace the missing values with None (Most likely the house has no masonry veneer)
train['Mas Vnr Type'] = train['Mas Vnr Type'].fillna('None')
train['Mas Vnr Type'].value_counts(dropna=False)
```

Out[18]:
```
None       1240
BrkFace     630
Stone       166
BrkCmn       13
Name: Mas Vnr Type, dtype: int64
```

Step 2: Replace other null values, mostly by None or 0.
(E.g. Residential area of null values for basement most likely because it has no basement.)

Data Cleaning - Ordinal or Nominal data

```
Pool QC (Ordinal): Pool quality

       Ex        Excellent
       Gd        Good
       TA        Average/Typical
       Fa        Fair
       NA        No Pool

Fence (Ordinal): Fence quality

       GdPrv     Good Privacy
       MnPrv     Minimum Privacy
       GdWo      Good Wood
       MnWw      Minimum Wood/Wire
       NA        No Fence

Misc Feature (Nominal): Miscellaneous feature not covered in other categories

       Elev      Elevator
       Gar2      2nd Garage (if not described in garage section)
       Othr      Other
       Shed      Shed (over 100 SF)
       TenC      Tennis Court
       NA        None
```

Step 3: From the data dictionary, check if a data is Ordinal or Nominal.

Ordinal data - Mapping

Mapping the `Lot Shape` data.

In [98]:
```
train.loc[:,'Lot Shape'].value_counts(dropna=False)
```

Out[98]:
```
Reg     1295
IR1      691
IR2       55
IR3        8
Name: Lot Shape, dtype: int64
```

In [99]:
```
Lot_dict = {'Reg':1 , 'IR1':2 , 'IR2':3 , 'IR3':4}
```

In [100]:
```
train['Lot Shape'] = train['Lot Shape'].map(Lot_dict)
```

In [101]:
```
train.loc[:,'Lot Shape'].value_counts(dropna=False)
```

Out[101]:
```
1    1295
2     691
3      55
4       8
Name: Lot Shape, dtype: int64
```

Step 4: Mapping of all Ordinal data.

Ordinal data - One-Hot Encoding

Lot hape	Utilities	Land Slope	Overall Qual	Overall Cond	...	Sale Type_COD	Sale Type_CWD	Sale Type_Con	Sale Type_ConLD	Sale Type_ConLI	Sale Type_ConLw	Sale Type_New	Sale Type_Oth	Sale Type_VWD	Sale Type_WD
1	1	1	6	8	...	0	0	0	0	0	0	0	0	0	1
2	1	1	5	4	...	0	0	0	0	0	0	0	0	0	1
2	1	1	7	5	...	0	0	0	0	0	0	1	0	0	0
1	1	1	5	6	...	0	0	0	0	0	0	0	0	0	1
2	1	1	6	5	...	0	0	0	0	0	0	0	0	0	1

Step 5 : One-Hot Encode all the nominal data.

Feature Selection & Data Modelling

The Curse of High Dimensionality

- By moving from 81 to 200+ features, **we've greatly increased the dimensionality of our data.**

- Features not truly associated with our target will create **noise**, which will lead to a deterioration in the model.

- **This increases the risk of overfitting**, as noise features may be assigned nonzero coefficients due to chance associations with the target variable.

1. Pairwise Correlation Analysis

1. **Identify Pairs of Highly Correlated Variables**

2. **Drop or Combine Variables**

Drop

v1	v2	pair_corr
Central Air_N	Central Air_Y	1.000000
Bldg Type_Duplex	MS SubClass_90	1.000000
Street_Grvl	Street_Pave	1.000000
Exterior 1st_CemntBd	Exterior 2nd_CmentBd	0.988254
Bldg Type_2fmCon	MS SubClass_190	0.977762
Exterior 1st_VinylSd	Exterior 2nd_VinylSd	0.977557
Exterior 1st_MetalSd	Exterior 2nd_MetalSd	0.976456

Combine

2. Variance Analysis

1. **Drop variables with below a variance threshold (e.g 99.5% single value)**

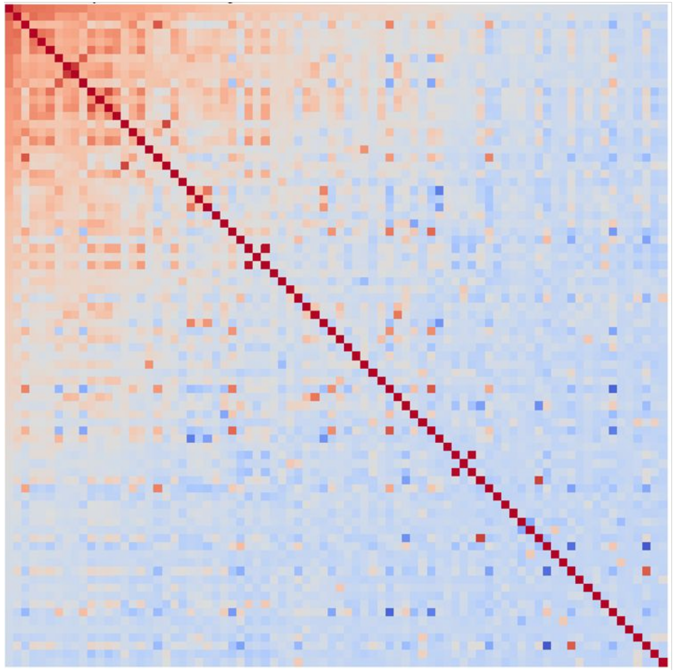
Neighborhood_Landmrk	0.000488
Condition 2_RRAn	0.000488
MS SubClass_150	0.000488
Condition 2_RRAe	0.000488
MS Zoning_I (all)	0.000488
ExtImStucc	0.000488
Roof Matl_Membran	0.000488
ExtStone	0.000488
Misc Feature_TenC	0.000488
ExtCBlock	0.000488

Drop

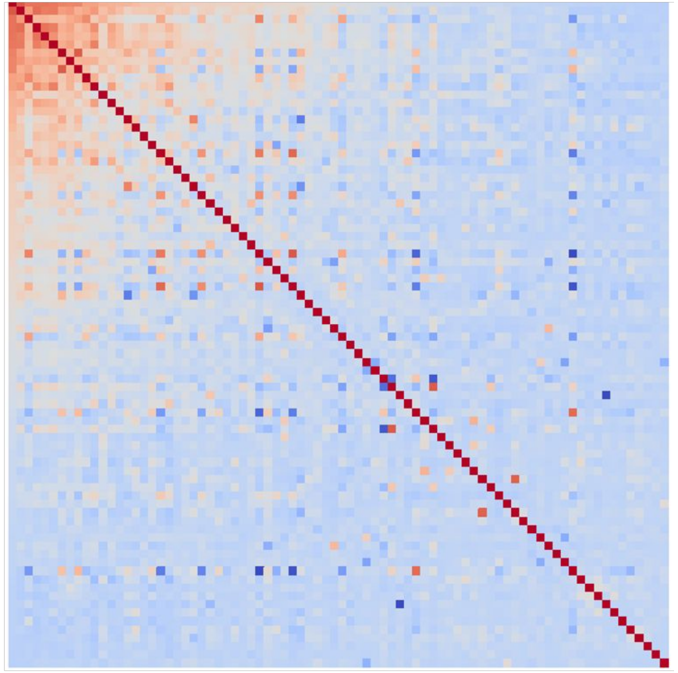
```
# Dropping features with low variance (<0.009)
low_var_drop_list = [item for item in low_var_list.index]
housing = housing.drop(low_var_drop_list, axis=1)
```

The Effects of Dimensionality Reduction

Before Reduction



After Reduction



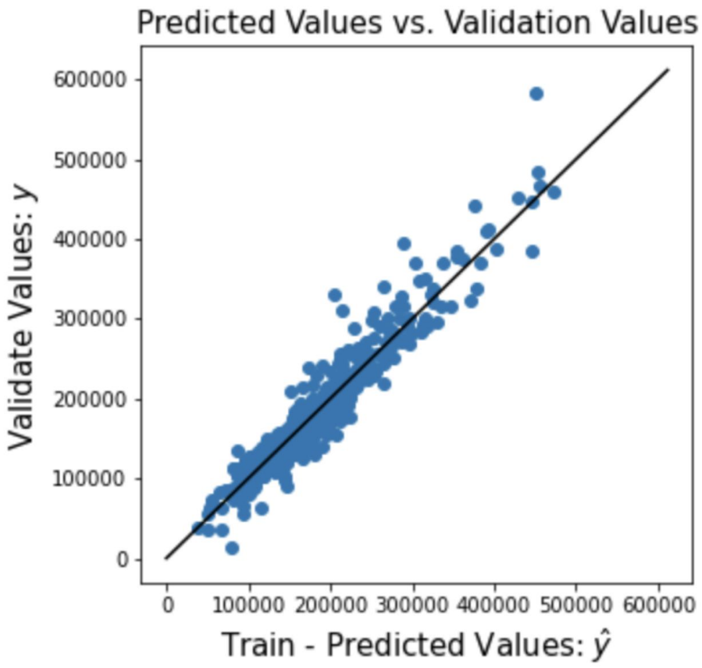
Model Selection

- As there's still a degree of multicollinearity in the data, we can **use regularization to further narrow down the total number of features**.
- Using techniques such as LassoCV, we were able to able to 'zero' out about 25 - 30 additional features features.
- A key question to consider is: **how features should your final model use?**
- This depends on the extent that you're willing to trade off interpretability for accuracy. With a higher number of features, we can gain higher levels of accuracy.
- A model with more features may be more accurate, but may have some limitations where **the predictors become difficult to interpret without extensive domain knowledge**.
- Ultimately, we prioritized interpretability and settled on a ridge regression model with 30 features.

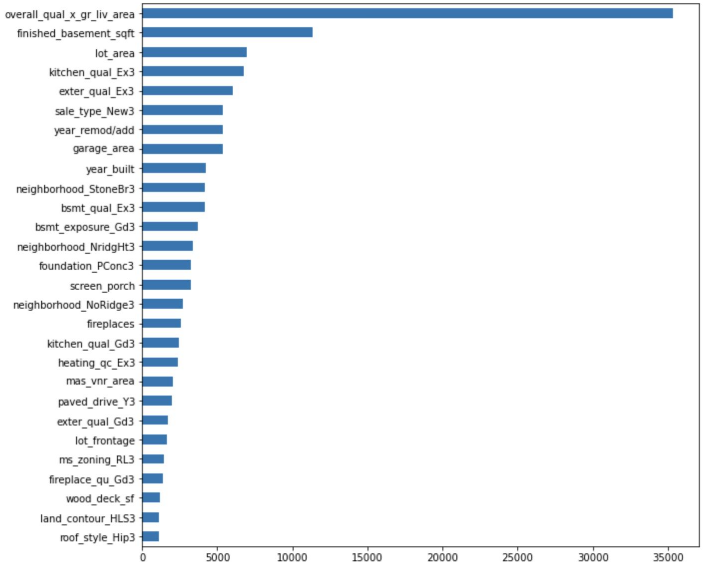
Conclusions and Recommendations

How good is our final model?

- Using the final model, we are able to **account for approximately 91% of the variation in Sale Price of a property** and is able to **predict the Sales Price within $23,000**. However, it is less accurate at predicting higher values.
- Caveats and areas of improvement:
  - Limited to Ames, and may not be generalisable to other cities.
  - 2006-2010 is during US subprime crisis, causing property price fluctuations
  - A more robust dataset with buyer demographic information could possibly help us segment buyers to provide more targeted recommendations.



How can we make this information useful for our target audience?



To make it more actionable for home sellers, I will lump these features into groups

Group	Consists of	Combined Impact
Interaction	• Overall quality + living area	$35,000
Area	• Basement sq footage • Lot area • Garage area • Wood deck sq footage	$27,000
Quality rating	• Kitchen quality • Exterior quality • Basement Quality • Fireplace quality	$25,000
Location	• Residential low-density zone • Northridge • Northridge heights • Stonebrook • Land contour - hillside	$12,000
Home age	• Year of remodelling • Year built	$9,000
Additions	• Fireplaces • Roof style • Central Air	$5,000

How can this analysis be used to inform seller decisions?

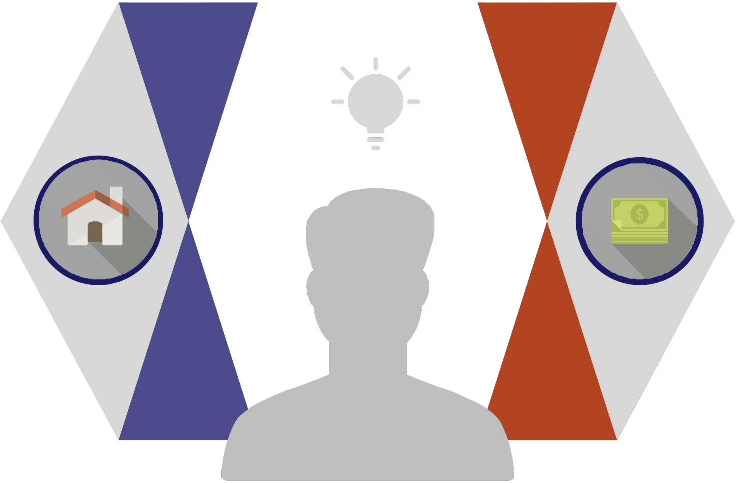
Given that it will be unlikely or extremely difficult to increase any continuous variables (such as lot frontage or square footage), we have decided to base recommendations on 2 groups of categorical variables as these can be changed by sellers.

Quality Ratings

Installation of new fixtures and fittings could lead to an increase in quality ratings, eg-

Having excellent kitchen quality will result in **$6773 increase in sale price**

Having excellent exterior quality will result in **$6055 increase in sale price**



Home Additions

Adding new features to your home can also drive up sale price, eg-

Having a paved drive will result in **$1964 increase in sale price**

Having a hip style roof will result in **$1107 increase in sale price**

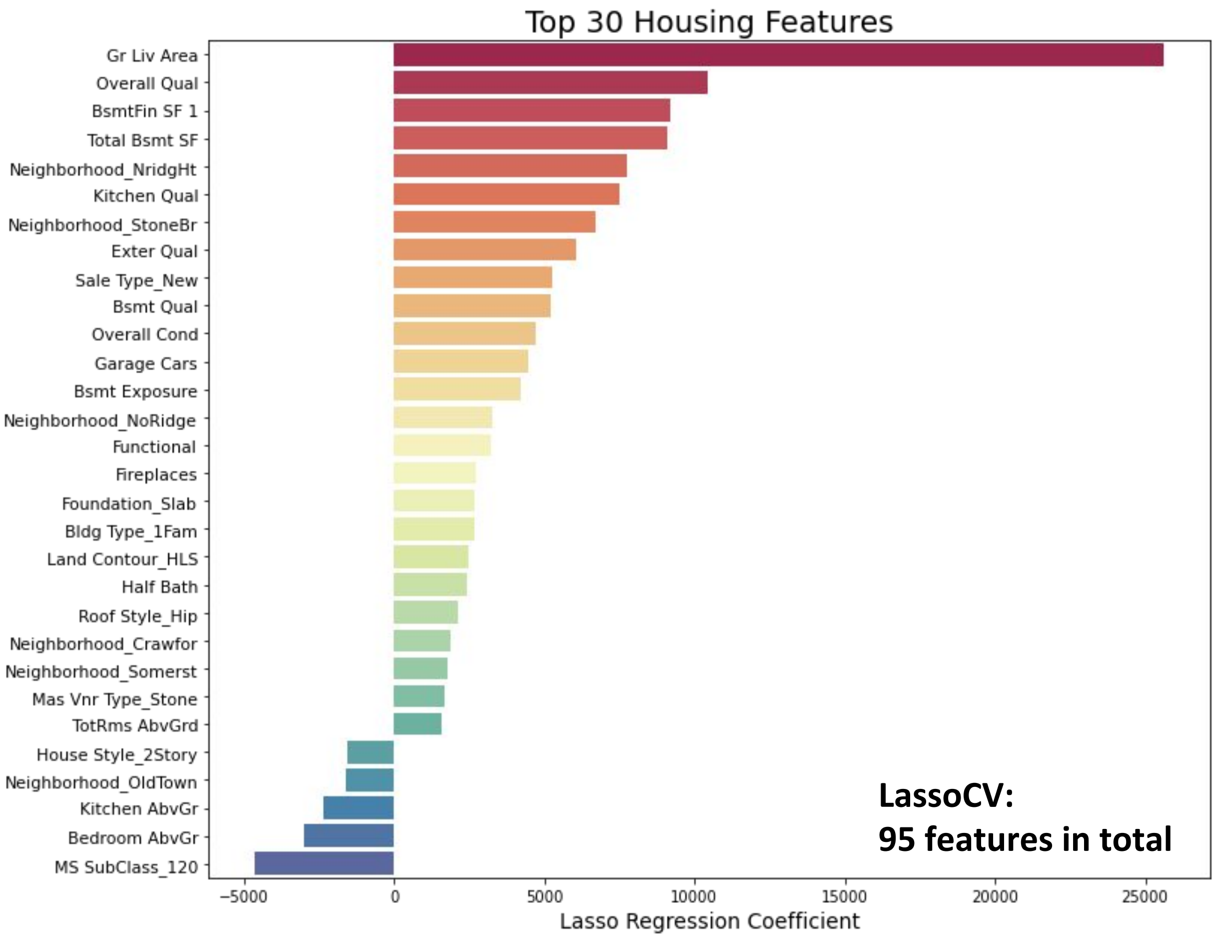
Appendix

Overview of data

- 2050 records of home sales in Ames, Iowa from 2006 to 2010.
- Data contains 80 ‘features’/variables including, but not limited to -
  - (Categorical) Type of housing/sale
  - (Continuous) Year of sale/remodelling/construction
  - (Continuous) Square footage of houses/bedrooms/garage/basement
  - (Ordinal) Rating quality of overall house/kitchen/basement/heating etc.
- **Our target variable for this analysis is to derive sales price.**

Comparison of regression models (post regularization)

Model	Train RMSE	Validate RMSE
Linear Regression	22588	22565
Lasso	22589	22564
Ridge	22591	22582
ElasticNet	23888	24056



For Kitchen AbvGr, I realized that houses with two kitchens have a lower mean sale price and were older compared to houses with one kitchen. In Iowa, summer kitchens were used prior to electricity and air conditioning to keep the heat from cooking out of the house during hot summer months.

In colder months, the indoor kitchens was used to help keep the house warm. The fact that houses with two kitchens generally don't have much porch square footage supports this idea (as summer kitchens are generally located on the back porch). This suggests that houses with two kitchens are more likely to be antiquated houses without a good heating/ventilation system.