



Project 2: Ames Housing Prices

Wee Zi Jian

**PROBLEM
STATEMENT**

01

METHODOLOGY

02

**DATA CLEANING /
EXPLORATORY DATA
ANALYSIS**

03

TABLE OF CONTENTS

04

MODEL BENCHMARK

05

**MODEL TUNING /
PRODUCTION**

06

CONCLUSION

An abstract geometric pattern consisting of white lines and dots (nodes) connected to form a network of triangles and polygons, set against a teal background. The pattern is more dense in the upper right and lower right areas, with some isolated nodes and small clusters in the upper right.

01

PROBLEM STATEMENT

SCENARIO

Prospective Ames home buyers / sellers need an estimate of **house prices** given the features of the house.



(<https://www.homebuilderdigest.com/best-custom-home-builders-in-iowa/>)

TASK

Create a Linear Regression model to accurately predict house prices, given Ames housing data.

Select 25-30 features out of 80, and refine model using cross validation and regularization.

Test the final Linear Regression model against unseen test data on Kaggle based on Root Mean Squared Error (RMSE) between predicted and actual sale prices.

The background is a solid teal color. Overlaid on this are several abstract geometric patterns. These consist of white dots (nodes) connected by thin white lines (edges). The connections form a complex, interconnected network of triangles and polygons. Some areas are more densely connected, while others are more sparse. The overall effect is a modern, digital, or scientific aesthetic.

02

METHODOLOGY

METHODOLOGY



DATA CLEANING / EDA

Clean and explore data to select first set of features for modelling



PRE-PROCESSING

Split original training data into sub train / test (holdout) sets

Scale and binarize features as necessary



MODELLING

Create Linear Regression model using first feature set

Regularize Linear Regression model using Ridge, Lasso and Elastic Net regression



EVALUATION

Score Linear Regression models based on RMSE of cross validation and prediction of holdout test dataset



FEATURE SELECTION

Analyse top coefficients of regression models to fine-tune feature engineering and create new feature sets



PRODUCTION

Evaluate models of subsequent feature sets

Select best performer for submission



03

**DATA CLEANING /
EXPLORATORY DATA
ANALYSIS**

DATA OVERVIEW

2051

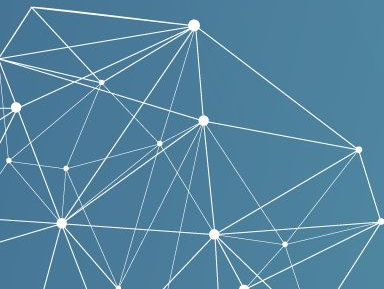
rows of housing data

80

columns of house features

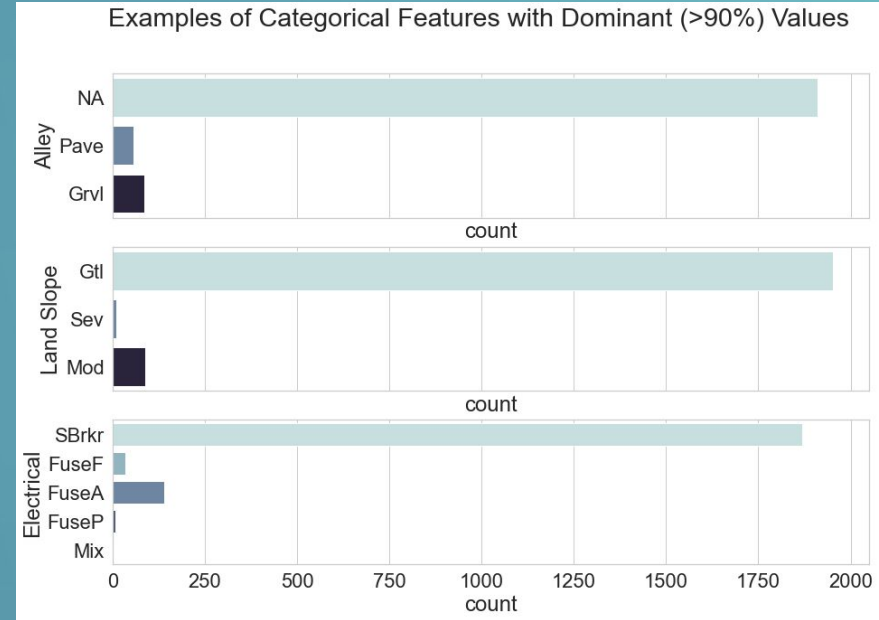
4

types of features (Nominal,
Ordinal, Discrete, Continuous)



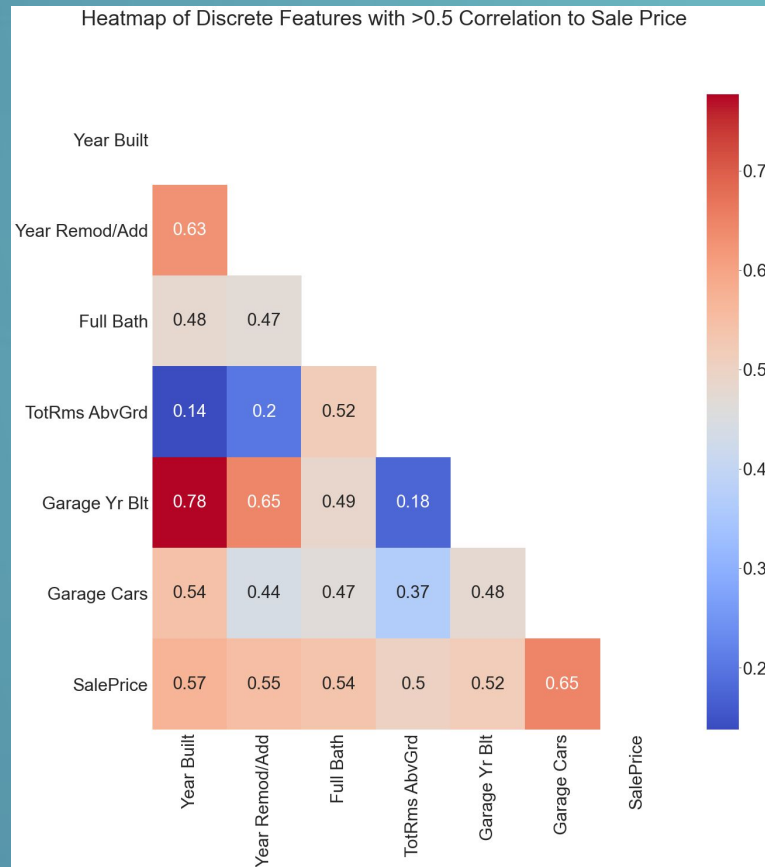
Categorical (Nominal + Ordinal) Features

- Fill in missing data as appropriate
- Check feature distributions
- **Remove features that have a predominant (>90%) value**
- Binarize remaining features
- Create empty columns to ensure train and test sets have equal number of columns



Numeric (Discrete + Continuous) Features

- Fill in missing data as appropriate
- Check feature correlation with Sale Price
- **Retain features with high (>0.5) correlation to Sale Price**
- Remove outliers
- Scale remaining features



First Feature Set

Nominal		Ordinal		Discrete	Continuous
MS SubClass	Roof Style	Lot Shape	BsmtFin Type 1	Year Built	Total Bsmt SF
MS Zoning	Exterior 1st	Overall Qual	BsmtFin Type 2	Year Remod/Add	1st Flr SF
Land Contour	Exterior 2nd	Overall Cond	Heating QC	Full Bath	Gr Liv Area
Lot Config	Mas Vnr Type	Exter Qual	Kitchen Qual	TotRms AbvGrd	Garage Area
Neighborhood	Foundation	Exter Cond	Fireplace Qu	Garage Yr Blt	
Condition 1	Garage Type	Bsmt Qual	Garage Finish	Garage Cars	
Bldg Type	Sale Type	Bsmt Cond	Garage Qual		
House Style		Bsmt Exposure	Fence		

04

MODEL BENCHMARK





Train / Test Split

20% of training data was kept as holdout set ($X_{\text{test}}, y_{\text{test}}$)

80% of training data was used for model training ($X_{\text{train}}, y_{\text{train}}$)

Benchmark RMSE

Testing the mean Sale Price of the training data (y_{train} mean) vs

the holdout set (y_{test}) gave a **RMSE score of 71,351**



Ordinary Linear Regression

Ordinary Linear Regression model with first feature set was fitted on training data (X_{train} , y_{train})

5 fold cross validation of training data scored a RMSE of $\sim 2.63 \times 10^{13}$

Testing against the holdout set (y_{test}) gave a RMSE score of $\sim 8.46 \times 10^{14}$

Regularization

5 fold cross validation of Ridge, Lasso and Elastic Net Regression was performed to reduce overfitting

New models generated the following RMSE scores:

	Cross Validation	Test Predictions
Ridge	19,950	58,395
Lasso	20,837	50,134
Elastic Net	20,837	50,134



05

MODEL TUNING /
PRODUCTION

Top 50 coefficients of Lasso Regression of first feature set were used for the second feature set

Feature	Coefficient	Coefficient
Gr Liv Area	24,095	24,095
Kitchen Qual_Ex	23,847	23,847
Bsmt Qual_Ex	18,516	18,516
Neighborhood_NridgHt	17,397	17,397
Bsmt Exposure_Gd	15,177	15,177
...
...
BsmtFin Type 1_LwQ	-1,100	1,100
Full Bath	-1,100	1,100
Heating QC_TA	-1,080	1,080
Lot Config_Corner	896	896
Exterior 2nd_Plywood	-802	802

Test Scores

		Cross Validation	Test Predictions
Benchmark	-	-	71,351
Feature Set 1	Ridge	19,950	58,395
	Lasso	20,837	50,134
	Elastic Net	20,837	50,134
Feature Set 2	Ridge	20,014	19,112
	Lasso	20,151	19,172
	Elastic Net	20,151	19,172

Subsequent feature sets were created based on top coefficients of preceding models



Feature Set 3


Top 30 coefficients of Lasso Regression of first feature set were used for the third feature set

Feature Set 4

Interaction features generated from third feature set were used for the fourth feature set

Feature Set 5

Top 30 interaction features from the fourth feature set were used for the fifth and final feature set



Test Scores

		Cross Validation	Test Predictions
Feature Set 3	Ridge	20,165	19,213
	Lasso	20,262	19,326
	Elastic Net	20,262	19,326
Feature Set 4	Ridge	18,581	18,173
	Lasso	18,783	18,472
	Elastic Net	18,783	18,472
Feature Set 5	Ridge	18,415	18,474
	Lasso	18,660	18,514
	Elastic Net	18,660	18,514

Final Feature Set

Feature	Coefficient	Coefficient
Gr Liv Area Bldg Type_1Fam	6,824	6,824
Overall Qual	6,656	6,656
Bsmt Exposure_Gd Total Bsmt SF	6,184	6,184
Gr Liv Area Foundation_PConc	4,608	4,608
Bldg Type_1Fam Overall Cond	4,236	4,236
...
...
Overall Qual Foundation_PConc	897	897
Overall Qual Overall Cond	747	747
Overall Qual Bldg Type_1Fam	-741	741
Overall Qual BsmtFin Type 1_GLQ	714	714
Total Bsmt SF Foundation_PConc	504	504

Production

Ridge Regression using **Feature Set 5** was the best performing model and thus selected as the production model

New Ridge Regression model using Feature Set 5 was created, optimized with cross validation, and fitted onto full training dataset

Final model generated cross validation and Kaggle test prediction scores as follows:

	Cross Validation	Test Predictions
Training data	18,449	
Kaggle Public Score		24,979
Kaggle Private Score		29,167



06

CONCLUSION

Model Analysis



Majority of the features in the final model are **positively correlated** with Sale Price.

Overall Qual and Overall Cond appear most frequently among the interaction features. Gr Liv Area and Total Bsmt SF also appear to be associated with high absolute coefficients.



A few features are **negatively correlated** with Sale Price.

For example Condition 1: Artery (adjacency to an arterial street), BsmtFin Type 1: Unf (unfinished basement), and Fireplace Qu: NA (no fireplace).



Given that interaction features were used, the **interpretability** of the model is reduced as the coefficients are tied to combinations of individual features.

For example, the interaction feature Gr Liv Area - Bldg Type_1Fam has a coefficient of 6824. This means: given that a building has Bldg Type: 1Fam, a unit increase with Gr Liv Area will be associated with an increase of 6824 in Sale Price. It is thus difficult to isolate the individual impact of each feature.

Recommendations



Buyers

Home buyers should pay attention to Gr Liv Area, Total Bsmt SF, Overall Qual and Overall Cond as these are highly correlated with Sale Price.

Home buyers should expect house prices to be higher if these features are higher in value.



Sellers

Home sellers who wish to maximise their selling prices should improve features positively correlated with Sale Price, if it is within their means, such as Overall Qual or Overall Cond.

Home sellers should also be aware of the negative qualities of their house (such as adjacency to an arterial street) which may reduce the price buyers are willing to pay.



(<https://www.homebuilderdigest.com/best-custom-home-builders-in-iowa/>)



THANK YOU



Credits: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.