Project 2: Ames Housing Prices

Quek Zhi Qiang, Wee Zi Jian

TABLE OF CONTENTS

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 regression model to predict house prices with lowest possible Root Mean Squared Error (RMSE), given Ames housing data.

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

Test the finalised model against unseen test data on Kaggle using the RMSE metric that would be provided by Kaggle.

BACKGROUND

- Sale price of residential property in Ames town, Iowa
  - 2006 to 2010 period
  - Compiled by Prof. Dean de Cock
  - 2051 training data
  - 879 test data

- 81 attributes
  - 79 quantity and quality attributes
  - 2 identification columns

CHALLENGES

- Mismatch in column names with data dictionary
    - Sale Condition not included
- Interpretation of columns
    - E.g. Lot Frontage
- Missing data and selecting imputation methods
    - Nominal vs Ordinal vs Discrete vs Continuous
    - E.g. missing data for lot frontage

02
METHODOLOGIES

WORKFLOW

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 regression model using first feature set

Regularize regression models with Ridge, Lasso and Elastic Net regression

EVALUATION

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

FEATURE SELECTION

Iteratively create new feature sets based on RMSE scores

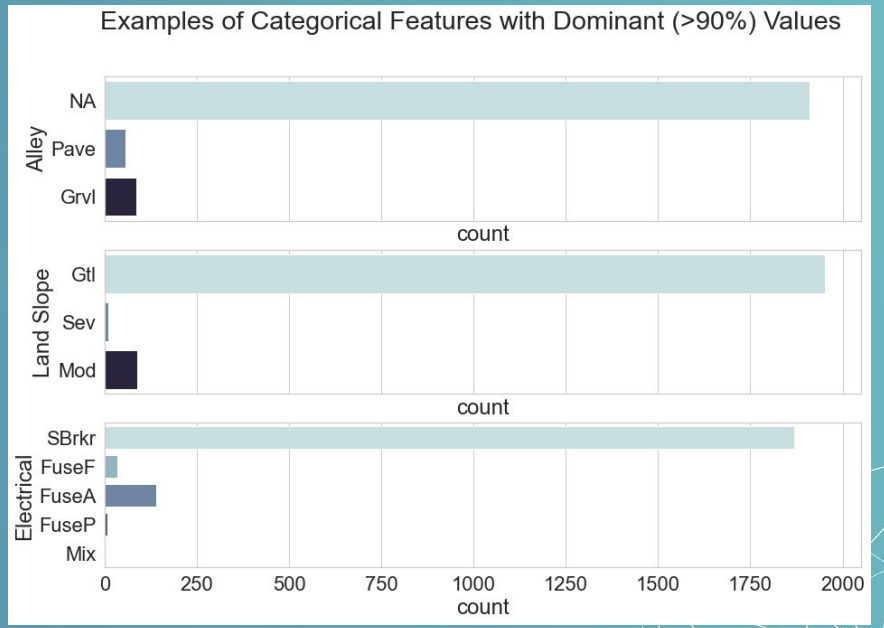
PRODUCTION

Evaluate models of subsequent feature sets

Select model with best test metrics for submission

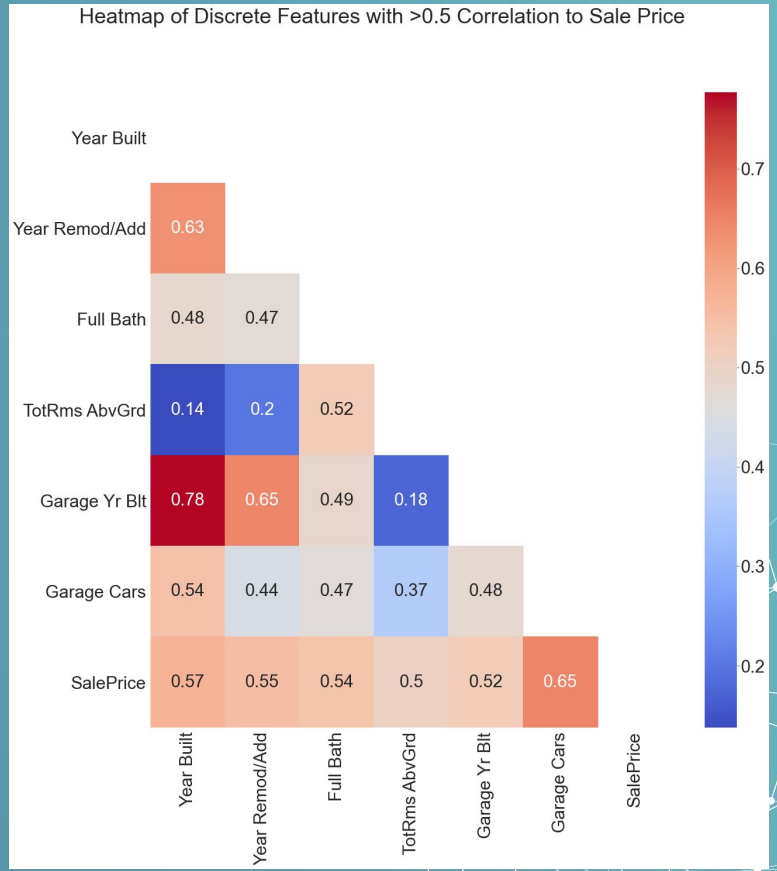
Categorical (Nominal + Ordinal) Features

- Fill in missing data as appropriate
- Check feature distributions
- **Remove features that have a predominant value (>90%)**
- 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

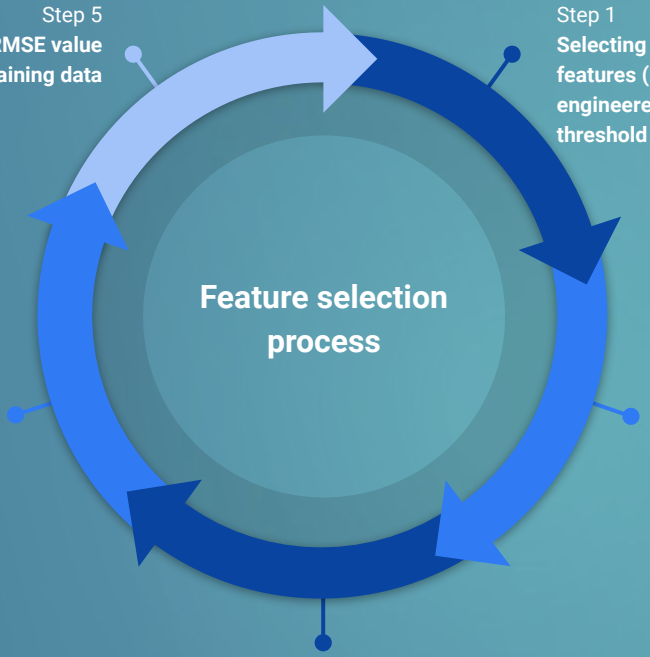


Iterative Feature Selection (Zi Jian)

Feature Set 1  Feature Set 5

1. Create model with first feature set after data cleaning / EDA
2. Select top Lasso coefficients of model to create subsequent feature set
3. Introduce interaction features
4. Select top Lasso coefficients of model with interaction features for subsequent feature set
5. Select best performing feature set / model for production

Iterative Feature Selection (Zhi Qiang)



03

RESULTS / CONCLUSION

Model Analysis (Zi Jian)

Ridge Regression using **Feature Set 5** was the best performing model and thus selected as the production model. It achieved a Kaggle public score of **24,979** and private score of **29,167**.

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

Overall Qual, Overall Cond, Gr Liv Area and Total Bsmt SF appear most frequently among the top interaction features.

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.

Model Analysis (Zhi Qiang)

Ridge Regression was the best performing model and thus selected as the production model. It achieved a Kaggle public score of **33,107** and private score of **34,118**.

Majority of the features in the final model are **positively correlated** with Sale Price. Out of the 40 variables (24 continuous and 16 dummified attributes from 3 nominal attributes)

- 15 negatively correlated
- 25 positively correlated

Top feature positively correlated to sale price: **Heating quality and condition** (25223)

Top feature negatively correlated to sale price: **Total porch area** (-21617, feature generated)

Interpretability was not affected as no interaction terms are involved

Feature engineered columns are mainly sum or division of columns

- Sum : Total porch area
- Division: Garage area per car

Recommendations

Buyers

Home buyers should pay attention to features which 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 within their means.

Home sellers should also be aware of the negative qualities of their house which may reduce the price buyers are willing to pay.

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

Potential Future Studies

Data

More granular data

Buyer and seller financial statuses and coordinates of the houses provide a better overview of transactions

Behavioural studies

Behaviours on seller and buyers

Human behaviours are found to have effects in transactions

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

THANK YOU