Project 2:
Ames Housing Sale Price Prediction

Presenters:

- Emily
- Ethan
- Gabriel
- Geoffrey
- Jin

Problem Statement

This project sets out to identify as accurate a model as possible for Ames housing sale prices based on the historical sales prices and corresponding information, using regression techniques, enhanced by applications of feature engineering, feature selection and regularisation.

Audience: for buyers and sellers of houses in Ames interesting in knowing the valuation of their house of interest

Data Sets

The dataset was prepared by Dean De Cock taken from the Ames, Iowa Assessor’s Office, originally used for tax assessment purpose. Data set contains information used in computing assessed values for individual residential properties sold in Ames, Iowa from 2006 to 2010.

The train data by Dean De Cook has 2051 observations and 82 columns which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables, and 2 additional observation identifiers. However, train data given by General Assembly has 81 columns, with 'Sale Condition' variable excluded, which include 22 nominal, 23 ordinal, 14 discrete, and 20 continuous variables, and 2 additional observation identifiers.

- nominal (categorical) meaning it takes qualitative values representing different categories
- ordinal(categorical) which provides an order of choices
- discrete(numerical) which are at set intervals
- continuous(numerical) which can take any value (such as square feet)

Assumptions and Limitation

Based on multi-linear regression’s assumptions:

- The predictors and target variable have an approximate linear relationship.
- Residuals are independent of each other, following a Normal distribution with mean 0 and have roughly equal variances.
- The predictors are independent of each other.

In addition, there are limitations to predict future prices based on a 2006-2010 dataset. The predictors for sale price may no longer be significant/relevant. Example, neighbourhood or street may have expanded.

Process

- Imputation of Missing Values
- Exploratory Data Analysis
  - Remove outliers
  - Correlation
- Feature Engineering:
  - Interaction terms
  - Feature Creation
  - Feature Selection
- One Hot Encoding & Label Encoding
- Model Prep
- Select Hyperparameters
- Scaling
- Baseline model
- Model Selection, Fitting and Evaluation
- Business Recommendation

Imputation of Missing Values

Column	Missing Value(s)	Type	Method
Total Bsmt SF	1	Numerical	Mean
Garage Cars	1	Categorical	Mode
Lot Frontage	330	Numerical	Linear
Garage yr built	114	Date	Yr Built Date

One Hot Encoding(OHE) & Label Encoding

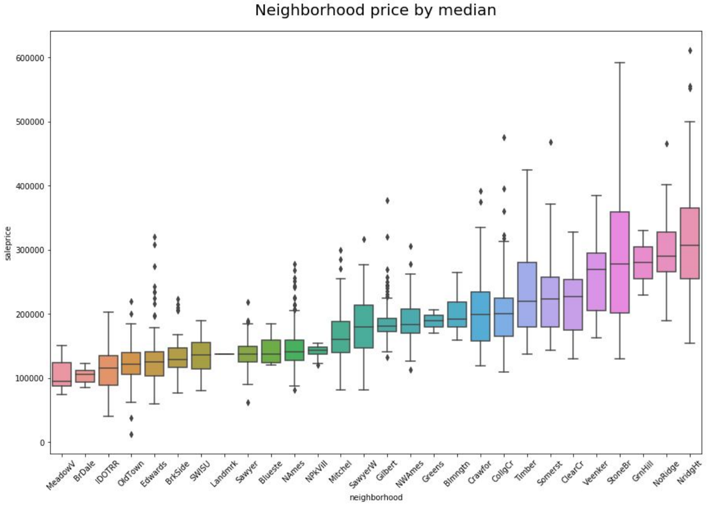
- Mapping of ordinal variables to numerical categories
- One-hot encoding of nominal variables via get_dummies

OHE (nominal)
Neighborhood
Ms Zoning

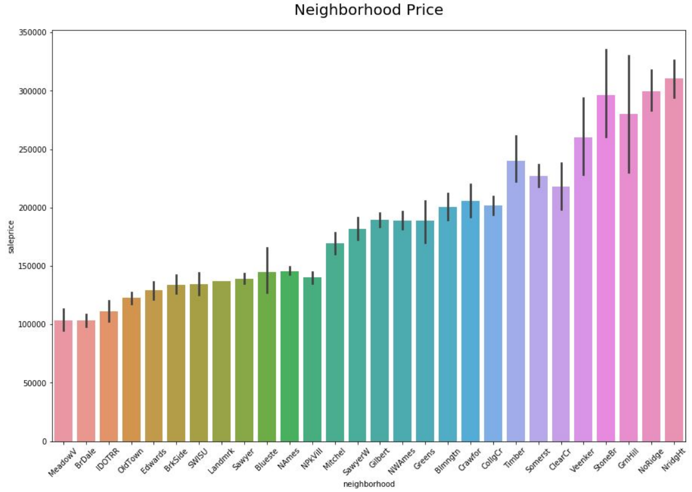
Label Encoding (Ordinal)		
Qual (Text)		Qual (Numerical)
None	“NA”	0
Poor	“Po”	1
Fair	“Fa”	2
Average	“TA”	3
Good	“Gd”	4
Excellent	“Ex”	5

One-Hot Encoding

- We can use boxplots to aid in selecting categories to encode
- Look at the pattern of boxplots and median sale price
- For example, we can see trending higher median sale price and perhaps some neighborhoods are more favorable than other



Neighborhood price by median



Neighborhood Price

Removing Outliers

Removing abnormal sales will improve prediction for sale prices. The abnormal sales can arise from exceptional high value sales or unusually areas size. Extreme outliers should therefore be removed. Let’s take a look at the following variables vs Sale prices:

- Lot Area
- Garage Area
- Pool Area
- Gr Liv Area
- 1st Flr SF
- Low Qual Fin SF
- Sale price

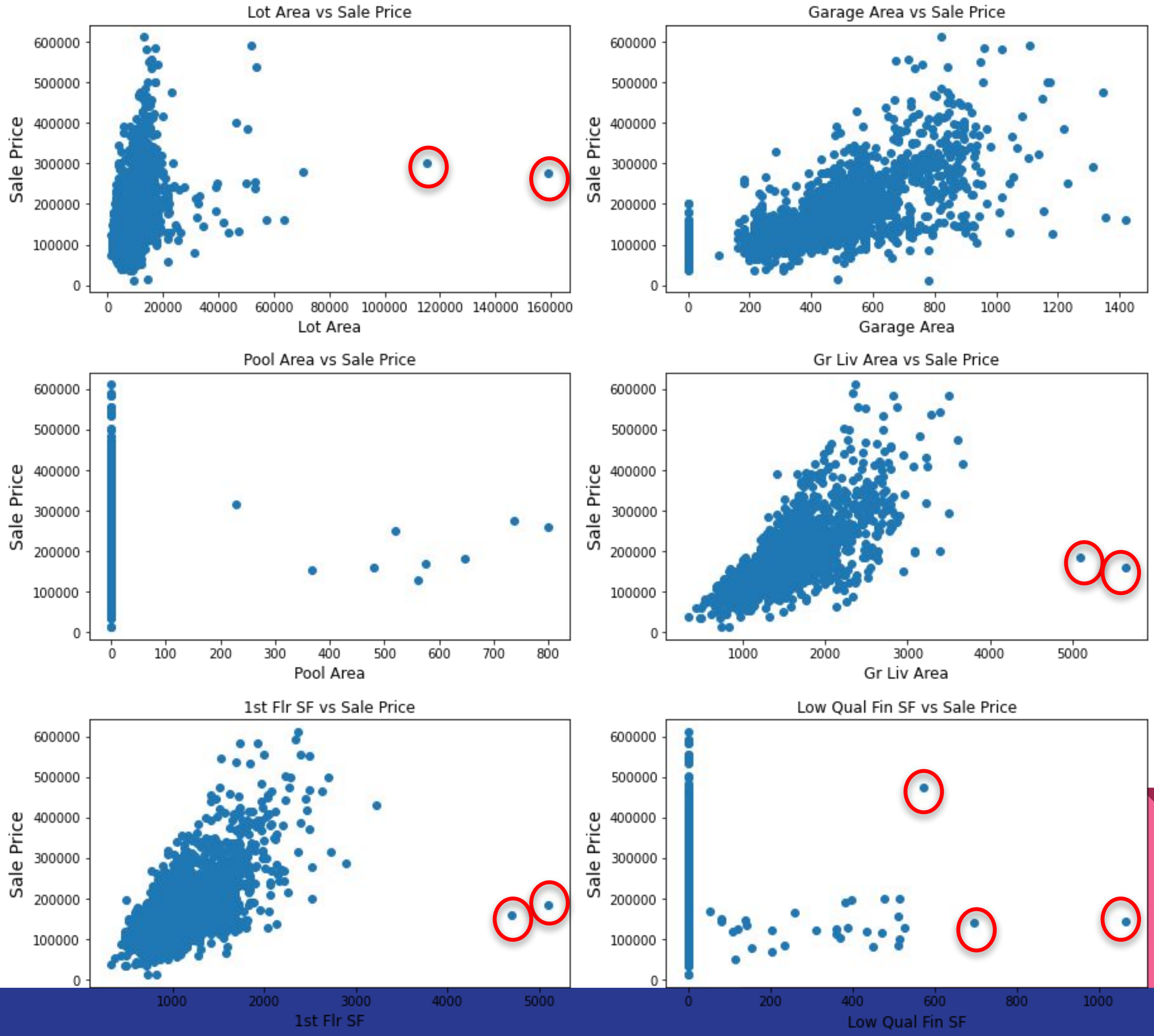
Removing Outliers

Notice 2 obvious outliers for '1st Flr SF', 'Gr Liv Area' and 'Lot Area'.

Notice 3 obvious outliers for 'Low Qual Fin SF'.

Notice only 9 non-zero 'Pool Area'. Do not foresee it will be a main predictor.

'1st Fir SF', 'Gr Liv Area' and 'Garage Area' have some linear relationship with 'SalePrice', suggesting they could be important predictors.



Correlation

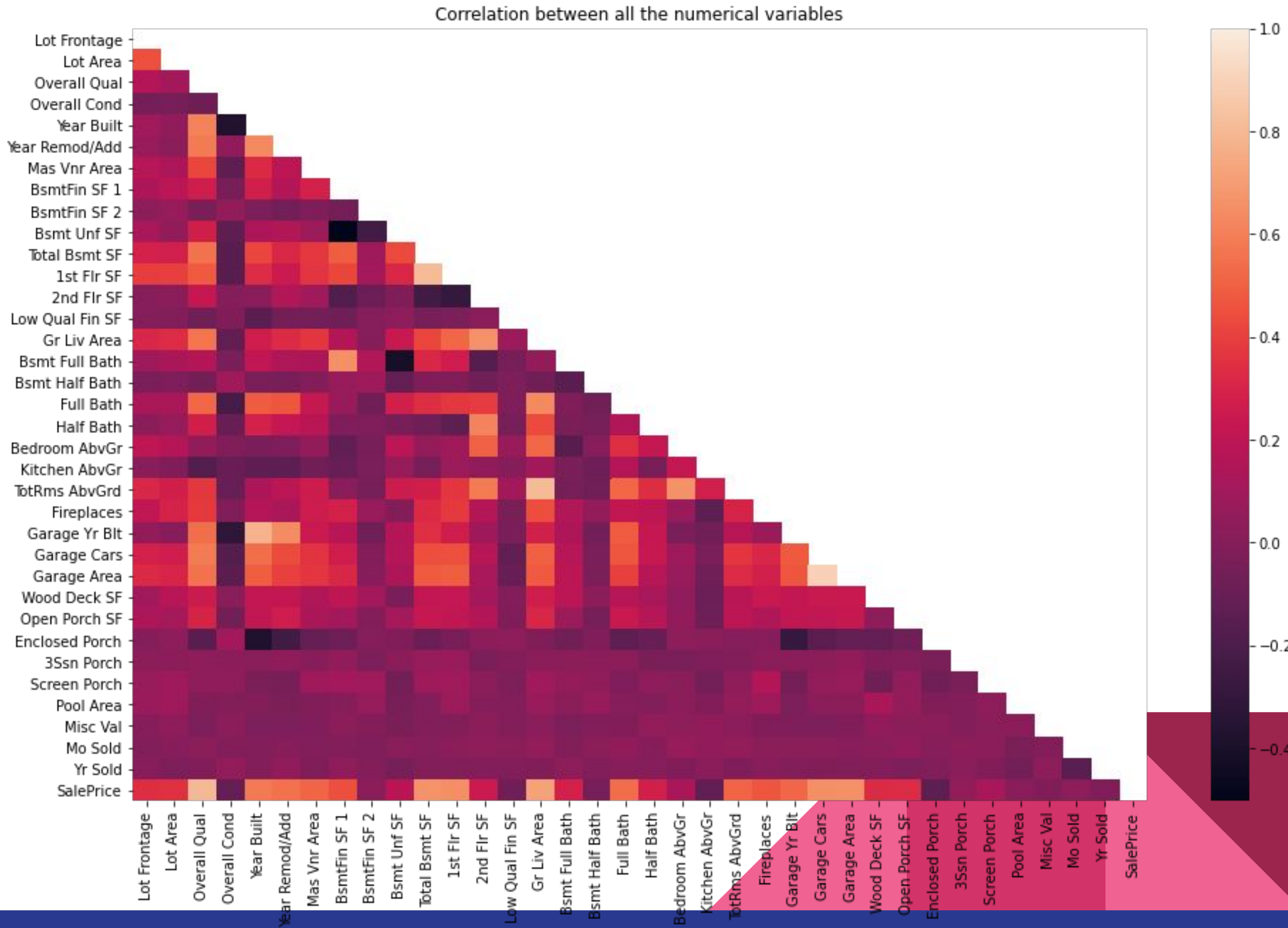
There are 6 obvious slight negatively correlated variables (~-0.4) such as:

- Overall Cond vs Year Built
- Overall Cond vs Garage Yr Blt
- Bsmt Full Bath vs Bsmt Unf SF
- Bsmt Unf SF vs BsmtFin SF 1
- Enclosed Porch vs Year Built
- Enclosed Porch vs Garage Yr Blt

There are 2 obvious highly positively correlated variables (~ 0.9) such as:

- TotRms AbvGrd vs Gr Liv Area
- Garage Area vs Garage Cars

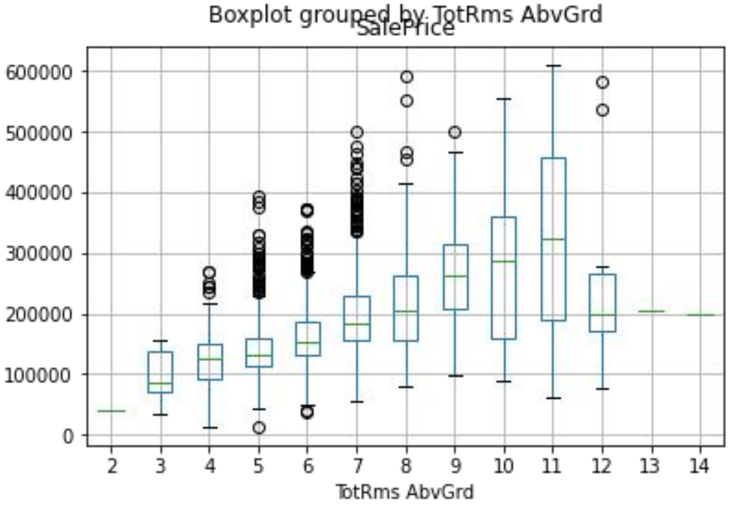
For these variables, **create interaction term for them**

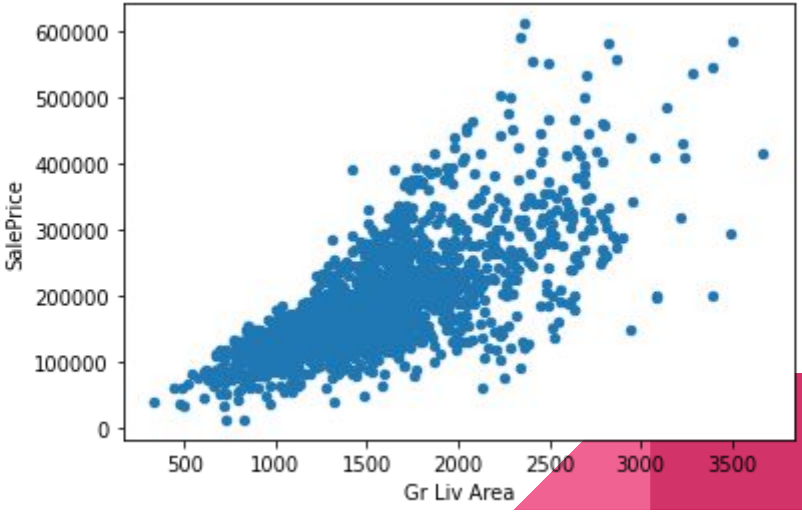


Correlation

For TotRms AbvGrd, Gr Liv Area, Garage Area and Garage Cars, I should drop two of them due to high correlation with each other, breaking the multi linear regression assumption. Earlier we discovered that Garage Area is likely a good predictor for SalePrice based on scatterplot. Based on above correlation heatmap, we can deduce that Garage Cars is of similar correlation coefficient with Garage Area to SalePrice.

Looking at the meaning, Garage Cars is size of garage in car capacity while Garage Area is size of garage in square feet. Hence, there is not much reason to keep both. Shall keep the more accurate measurement which is Garage Area and **drop Garage Cars**.





Both have convincing relationship with SalePrice. Looking at the correlation heatmap earlier, decide to **drop TotRms AbvGrd** instead.

Feature Engineering: Interaction terms

After investigating the numerical variables in terms of their qualitative meaning, I have selected these few that I suspect may be the main predictors to execute feature engineering:

Set 1:

- Overall Qual
- Overall Cond
- 1st Flr SF
- 2nd Flr SF
- Low Qual Fin SF
- Gr Liv Area

Set 2:

- Bsmt Full Bath
- Bsmt Half Bath
- Full Bath
- Half Bath
- Bedroom AbvGr
- Kitchen AbvGr

create higher polynomial of 2 and interaction terms for these

Feature Creation

- We could take difference of year sold and year built as the building age
    - df['building_age'] = df['yr_sold'] - df['year_built']

- Age correlation to sales = -0.58
- Year built = 0.58

Model Prep

The data set was further divided into two datasets, one of which SalePrice (training data) are provided to be used for modelling purposes, after which the selected features and models are used for the prediction of the sale prices on the test set

Test data given by Kaggle is only to be used for prediction. It comprises of the variables without Sale Price.

Select Hyperparameters

Using GridSearchCV

Use Lasso Regression to get alpha

Use Elastic Regression to get alpha and ratio

Refine Elastic Regression

Best hyperparameter for Elastic Reg: alpha = 0.1, ratio = 0.9

With R^2 = 0.94

Scaling

Perform StandardScaler

Purpose:

To transform the data in such a manner that converts the columns to Z-scores. Standardization is useful for data on different scales, separated by a few orders of magnitude (E.g. (square footage is in the thousands and number of bedrooms is in the single digits).

Baseline Model

Mean of SalePrice will be used as benchmark. If predictive model performs better, it shows improvement from baseline model which is simply taking the mean

Model Selection

Model	R^2 Accuracy of prediction (0 to 1)(higher better)	RMSE (lower better)
Baseline Model	0	30,642
Elastic Reg ✓	0.94	22,634
Lasso Reg	0.92	23,749

Model Fitting and Evaluation

Using Elastic Regression chosen earlier and fit the model with train(train.csv) data.

We estimate the R^2 and RMSE Score for test(train.csv) data.

Finally we fit the test.csv data given by Kaggle for prediction of SalePrice.

Business Recommendations

- We found that Overall Quality rating of the property is most influential when determining price, followed by the above ground living area and garage size.
- We found that the exterior quality was fair, that badly affected the value of the home and we could also see this trend with a similarly bad quality rating of the kitchen and basement.
- The neighborhoods that seem to command higher sale prices were Stone Brook, Northridge, Northridge Heights and Green hills.
- If this model were applied to other cities, the ordinal columns that reflect the quality and condition of the property should be standardised and using a similar scale. The neighborhoods that are present in the features should be dropped as it would be irrelevant to modelling other cities. There might be differing preferences of local buyers depending on the weather and demographics of the city which will impact what is a 'desirable' property.

Thank you!