DSI Project 2

What features affect the Sale Price of a house?

Presented by Yong Fah Aik

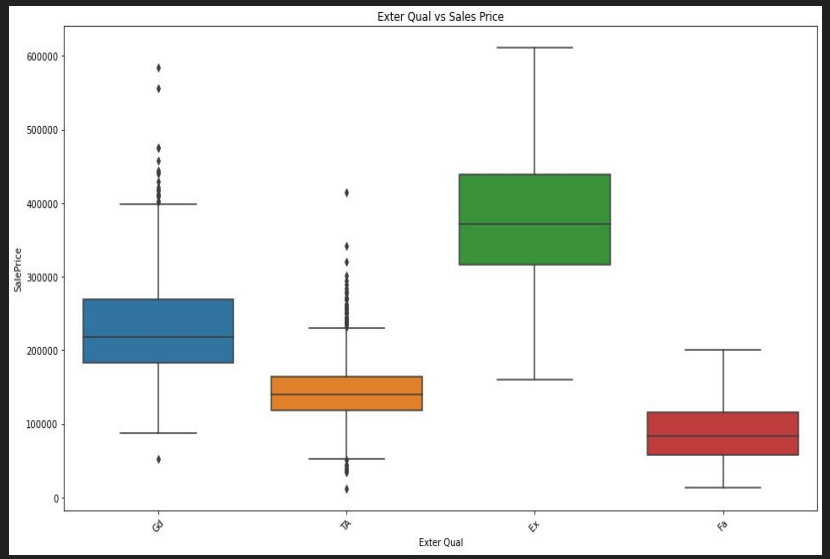
Table of Contents

Problem Statement

As a consultant to a potential house-owner in the city of Ames, Iowa, this project is tasked with the following:

- The identification of the key features that will affect the Sale Price of the house
- Recommendations on the features of the house that will affect the Sale Price the most, and so enable the potential house-owner on judging what sort of houses to look for based on budget.

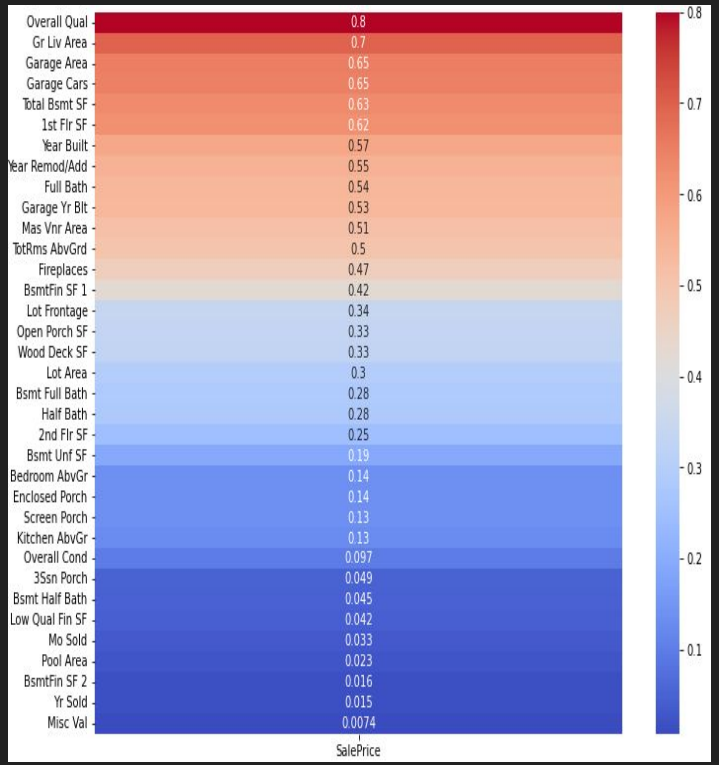
Methodology: Data Cleaning



Ordinal Column to be converted, with missing values of ‘NA’ that is imputed as zero

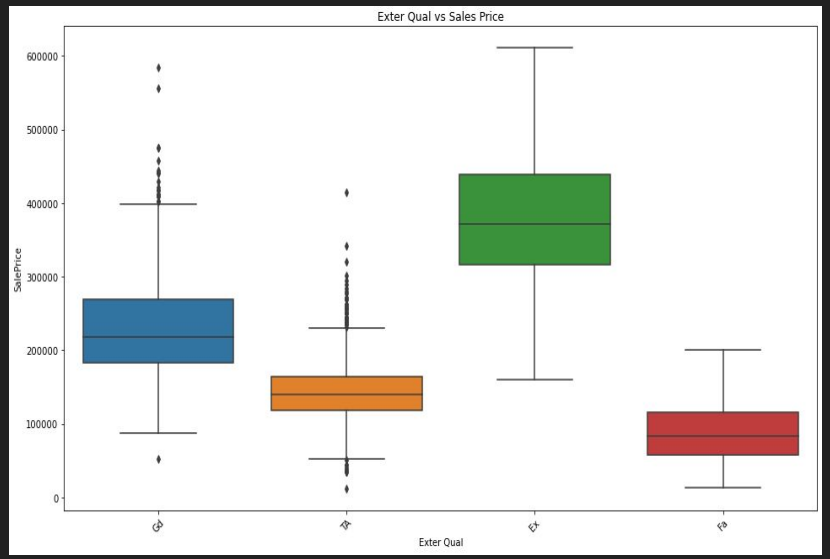
- Dropped columns with too many missing values
- Imputation of missing values with mean or zero values (depending on data type) for the remaining columns
- Conversion of ordinal columns to discrete columns
- Renamed columns for easier manipulation of data
- Removed outliers (ie ‘Gr Liv Area’ > 4000, ‘Lot Area’ > 40000)

Methodology: Feature Selection (Numerical)



- Features with high correlations to Sale Price are pre-selected, ie > 0.5
- Then, among those with high correlations, the other columns that are similar are dropped (ie dropping Garage Cars due to similarity to Garage Area)
- Features with lower correlations but seems important are chosen, ie Lot Frontage, Lot Area, Fireplaces

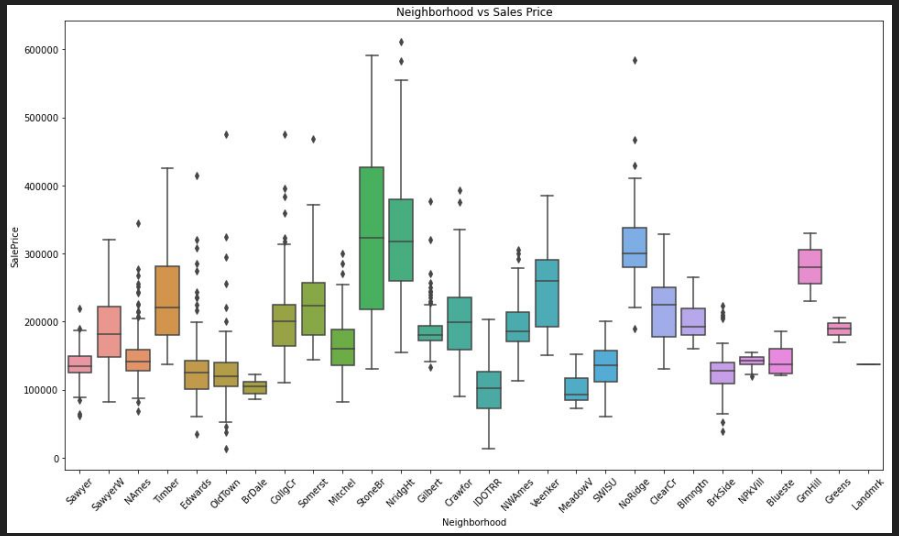
Methodology: Feature Selection (Categorical)



Ordinal Column to be converted, with missing values of ‘NA’ that is imputed as zero

- Features being mostly composed of a single value (mode) at above 85% are not considered at first pass
- Box plots against Sale Price are then drawn for the rest
- Ordinal Columns (like Exter Qual column on the left) with clear correlation to Sale Price are chosen.

Methodology: Feature Selection (Categorical) cont.



Categorical column of Neighborhood, where StoneBr, NridgHt and NoRidge are chosen as features

- Non-Ordinal Columns are then chosen based on whether there are clear correlations to Sale Price.
- For example, the Neighborhood feature on the left has 3 neighborhoods with a higher median Sale Price compared to others. As such, these 3 neighborhoods are dummified and chosen as features.

Key Findings (Model)

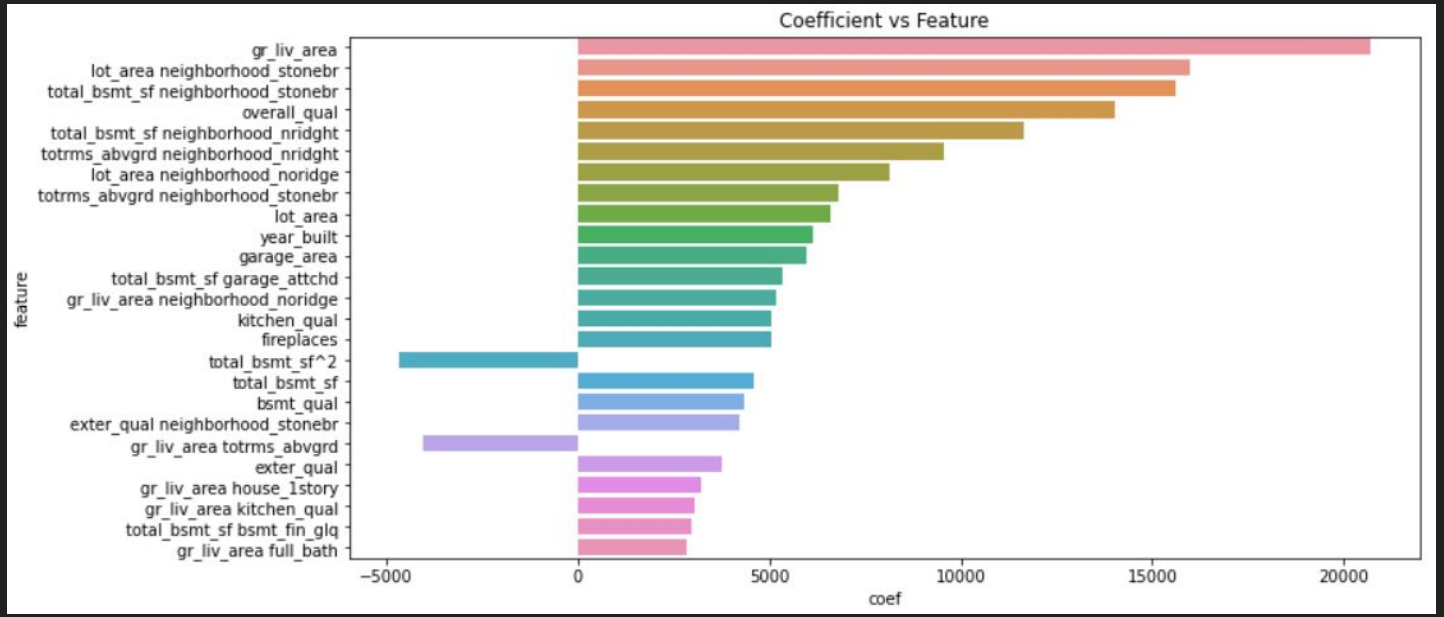
- Baseline made using the mean of the training data.
- Models based on OLS Linear Regression, Ridge Regression and Lasso Regression are then applied.
- Finally, interaction terms are added to the models.
- Evaluations are made using statistical metrics to show how well the model does.

Model	Train/Test RMSE	Train/Test MAE	Train/Test R2	Cross val RMSE	Cross val R2	No. of Features
Baseline	79526.8522 / 78375.2624	--	--	--	--	26
Linear Regression	30818.775 / 26866.9156	19916.61 / 19634.45	0.8498 / 0.8824	32499.3382	0.8313	26
Ridge Regression	30837.174 / 26679.0112	19898.7 / 19612.71	0.8496 / 0.8841	32500.1995	0.8315	26
Lasso Regression	30890.5437 / 26499.4763	19807.77 / 19494.47	0.8491 / 0.8856	32413.1619	0.8326	26
Lasso Regression (interaction terms)	28784.1962 / 25083.6057	18938.95 / 18487.04	0.869 / 0.8975	30763.6683	0.8484	32
Lasso Regression (polynominal features)	21398.005 / 22423.2328	14962.62 / 16136.06	0.9276 / 0.9181	27024.0471	0.8866	377

Key Findings (Final Model)

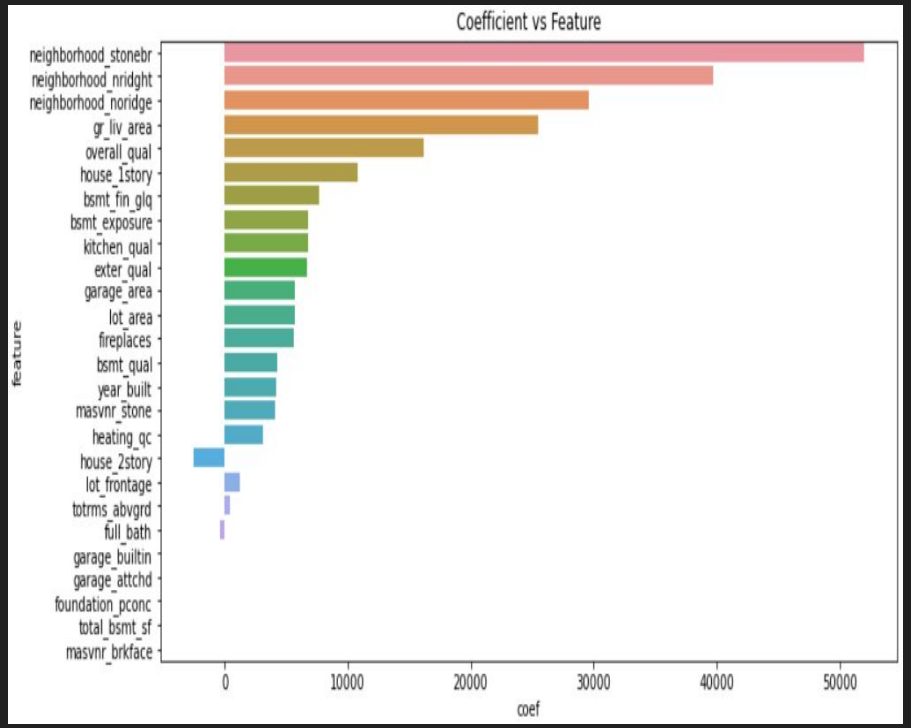
Plot of Actual Sale Price versus Predicted Sale Price for Final Model of Lasso Regression with Polynomial Features

Key Findings (Final Model)



Plot of coefficients of model vs feature, with the features arranged from the top based on impact of feature on model

Recommendations



Plot of coefficients vs feature for the Lasso Regression Model

Features that are important

- Neighborhood - Stone Brook, Northridge Heights, Northridge
- Overall Quality
- Living Area Above Ground
- House Style of 1 Story
- Lot Area
- Year Built
- Garage Area