Modelling for Ames Housing Data

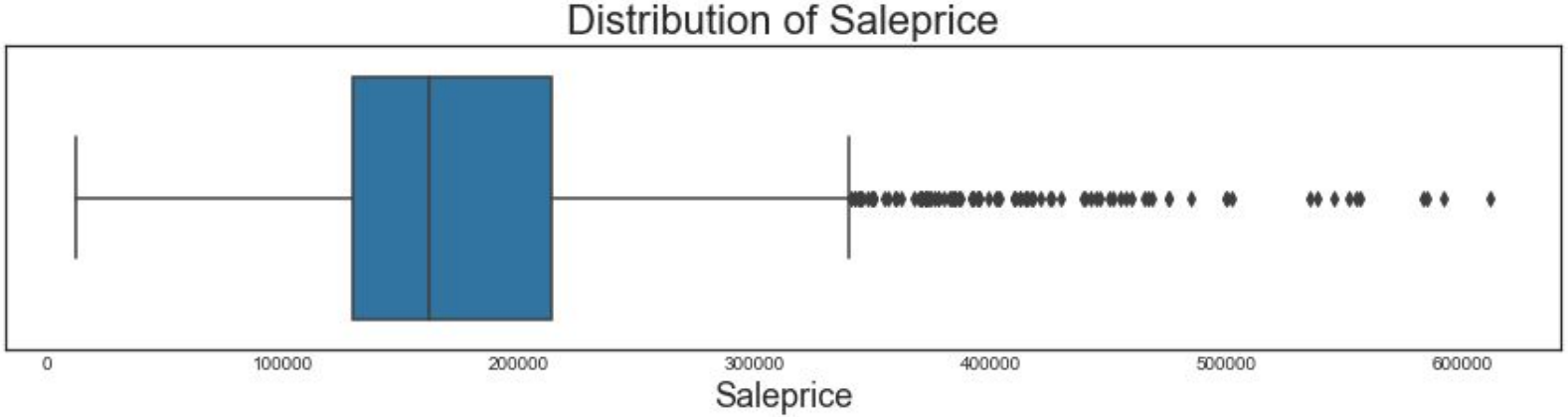
By Lin Junyuan

Problem Statement

What is the best regression model, with

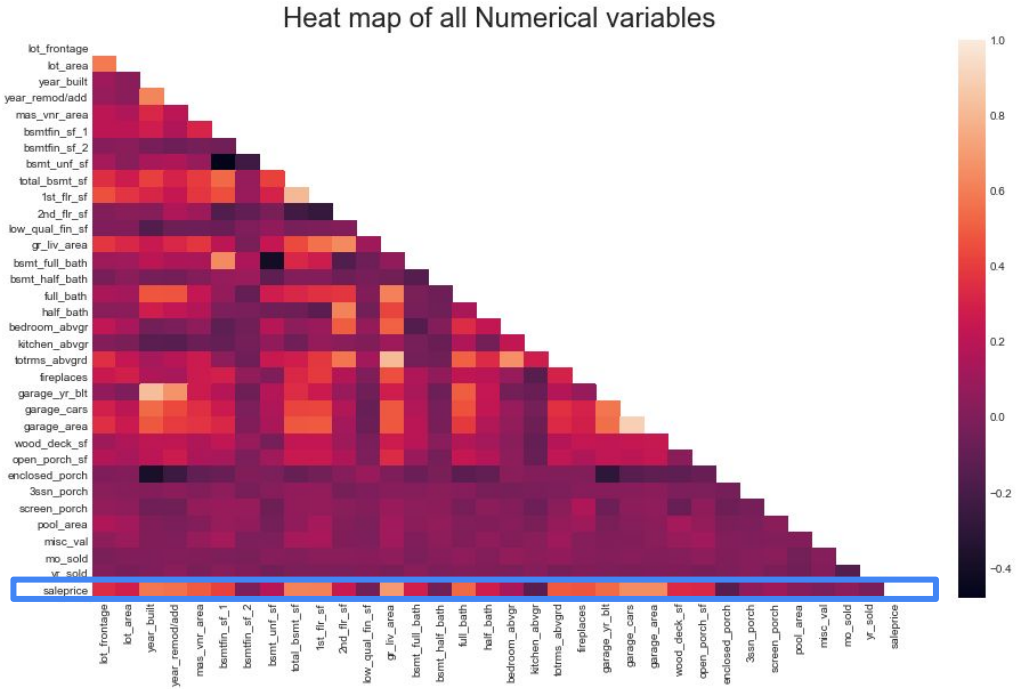
1. an upper limit of 30 variables, that can predict house sale price in Ames Iowa to
2. a RMSE of less than $40,000 and
3. what are its top 3 influencing factors.

EDA - Target variable : Saleprice



Distribution shows presence of outliers, which need to be removed to avoid skewing the model.

EDA - Numerical variables

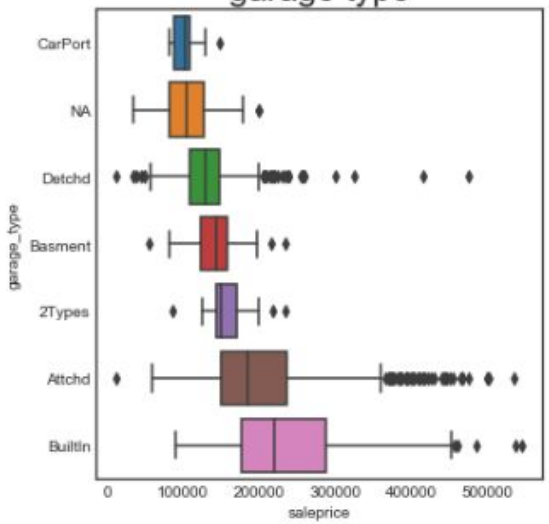


- Some variables are collinear
  - Needs to be identified and only one variable to be suitably select to represent the others
- A number of variables have moderate and strong correlation to saleprice

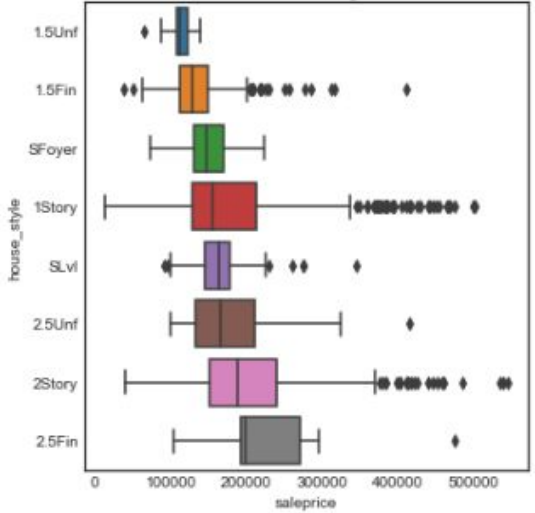
EDA - Numerical variables

- Boxplots against saleprice used to evaluate suitability
- Good distinct distribution between each unique values used as criteria.

G



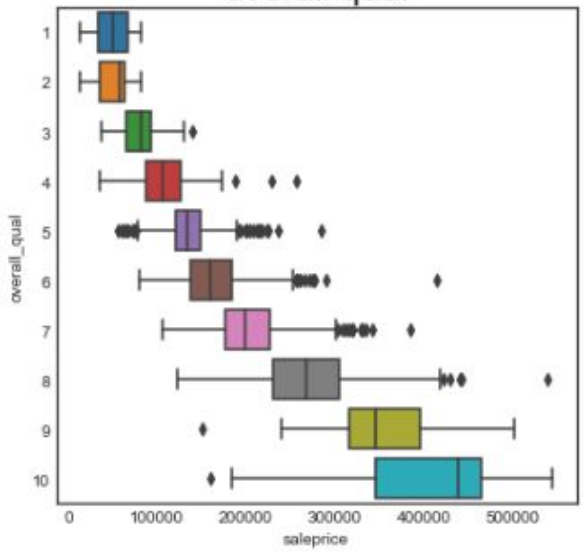
NG



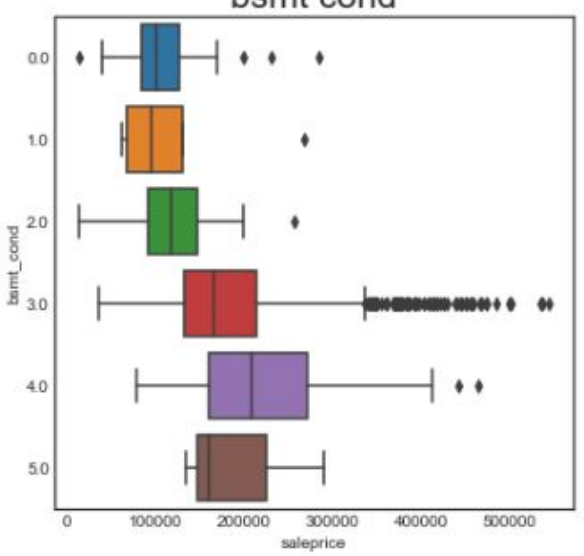
EDA - Ordinal variables

- Similar criteria to nominal variables
- Additional requirement for them to be monotonic
- Collinearity also needs to be assessed

G



NG



Feature Engineering

- These steps were implemented
  - Missing value Imputation
    - Mean for numerical variables, Mode for nominal and ordinal
  - One hot encoding for nominal variables
- Interaction terms were explored
  - However, viable pairs with high correlation to saleprice all contains variables which are already chosen in the previous steps.
  - No interaction terms were created to prevent collinearity

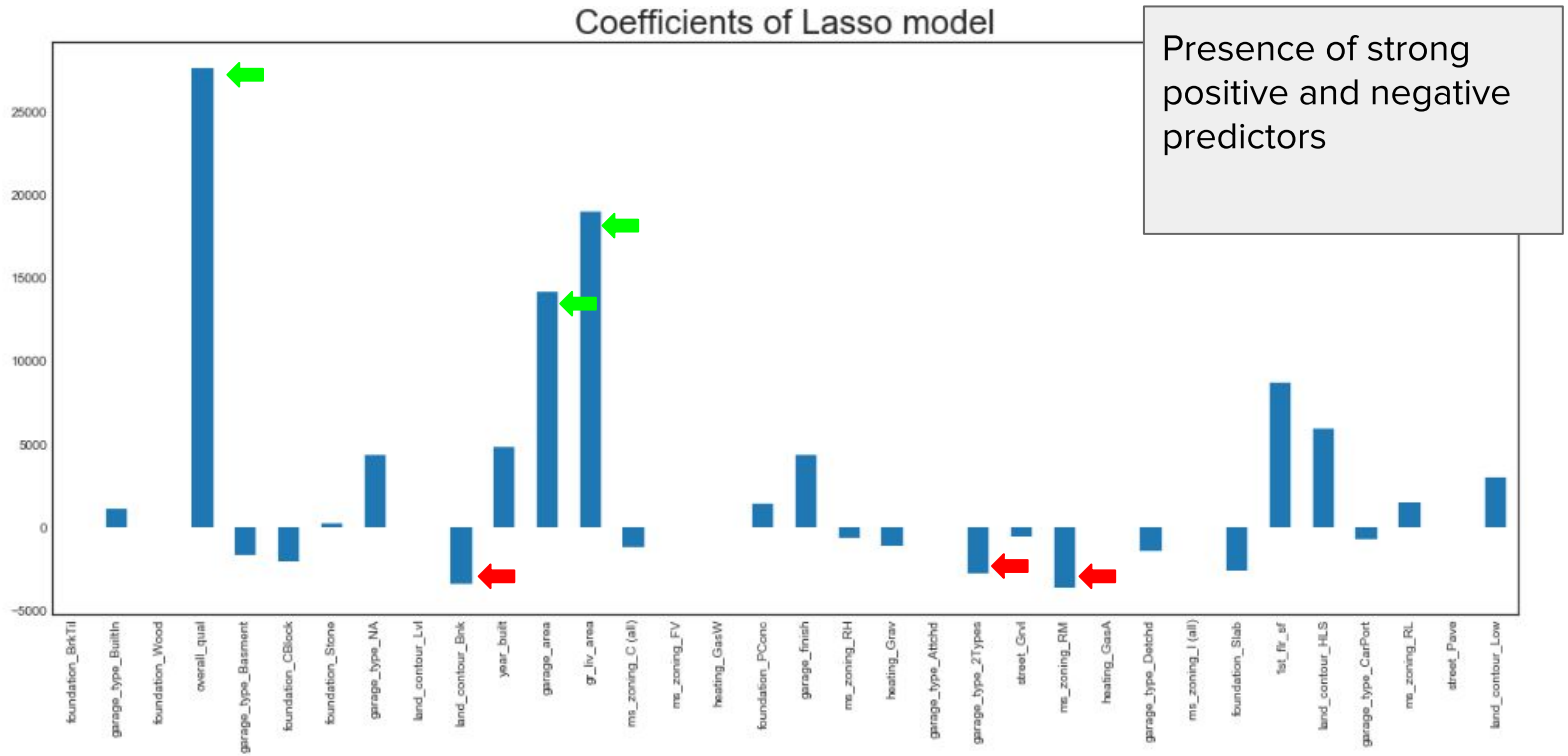
Modelling: Baseline

- Use mean of saleprice in training data as all predictions as baseline.
- Use of Root Mean Squared Error (RMSE) as metric as it is the one selected by the Kaggle Challenge
- RMSE: $79,435

Modelling: Baseline vs other models

Model	RMSE	No. of variables
Baseline	$79,435	NA
Lasso	$33,414	26
Ridge	$33,575	33
Elastic Net (Optimised using GridSearch)	$33,542	33

Best model: Lasso



Presence of strong positive and negative predictors

Top 3 positive predictors

1. Overall material and finish quality
2. Above grade (ground) living area in square feet
3. Size of garage in square feet

Top 3 negative predictors

1. Land Contour where there is a Quick and significant rise from street grade to building
2. There is more than one type of garage
3. Zoning classification of Residential Medium Density