DSI - Project 3

House Sales dataset (Ames, Iowa)

Fiona Dee - June 2021

Data: Preparation & Cleaning

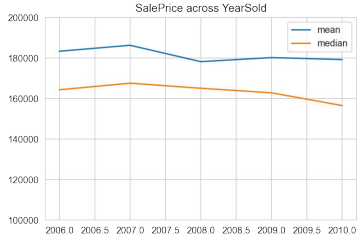
Start with information in data_description.txt

- Group into feature sets: land / building core / building other / sale.
- Identify type of each feature (and expected dtype): continuous / category / rating.

After loaded into dataframe:

- Remove non-residential (MSZoning=A,C,I). Keep FV (Floating Village Residential).
- Changed features with ratings to numbers (Ordinal Variables) e.g. Ex:5, Gd:4, TA:3, Fa:2, Po:1

Data: Decisions & Assumptions



Decisions:

- Identify extreme outliers & extremely low variance features
- Keep ‘NA’ where data description has it as a valid option

Assumptions

- There will be external economic factors which impact on SalePrice as well.
- Let the models find the signal in the noise (light touch cleaning & minimal imputing of missing values).

Fixed attributes: Feature Engineering

Decisions

- *Use features:*
  Land features /
  Building core features /
  Sales features
- Drop rows with extreme outliers & Utilities feature.

Assumptions

- Removing extreme outliers will reduce the variance for the models to work with.
- Leaving in less extreme outliers allows flexibility for models to choose features which I might dismiss.

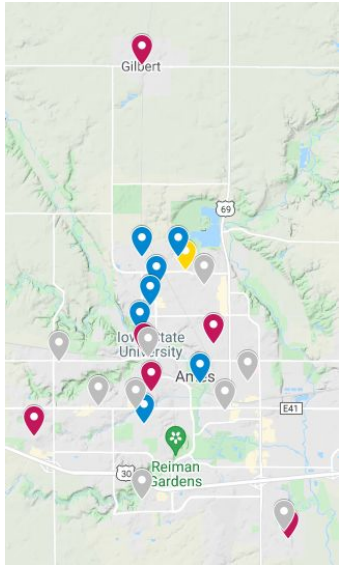
Fixed attributes: Model = Lasso

- Explains 89% of variance, with mean of ~ 4k (2% error for mean SalePrice)
- Residuals are reasonably balanced >$250k, although negative bias
- SalePrice has mean $181654, and median $163945

➢ Above Ground Living Area: 30,951

➢ Northridge Heights: 11,307

➢ Year Built: 10,002



Link to map

Renovate-able: Feature Engineering

Decisions

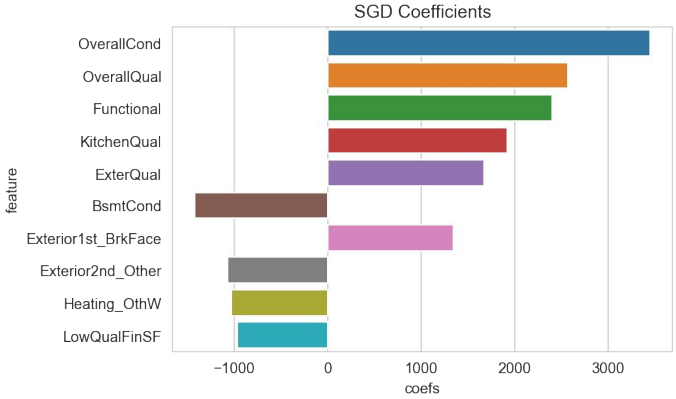
- *Use features:*
  Building other features /
  Sales features
- Match the rows dropped for extreme outliers with the fixed dataset.

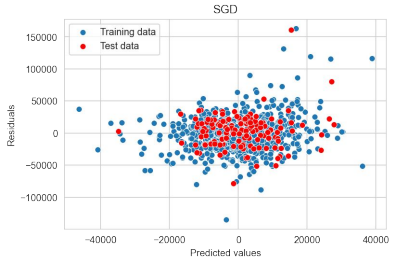
Assumptions

- Many quality features are subjective and they are likely to be highly influential on the model.

Renovate-able: Model = SGD (L2, Loss Sq)

- Explains ~14% of residual variance,
  with mean of ~ 3k
- Residuals are nicely centred around zero, although scatter increases beyond +/- 10K





Options for alternative analysis/prediction

- Change ordinal variables into dummies.
- Feature engineering: create new summary variables, e.g. Total sqft
- Impute rather than drop LotFrontage nulls
- Cull more outliers so there is less variability for models to deal with.