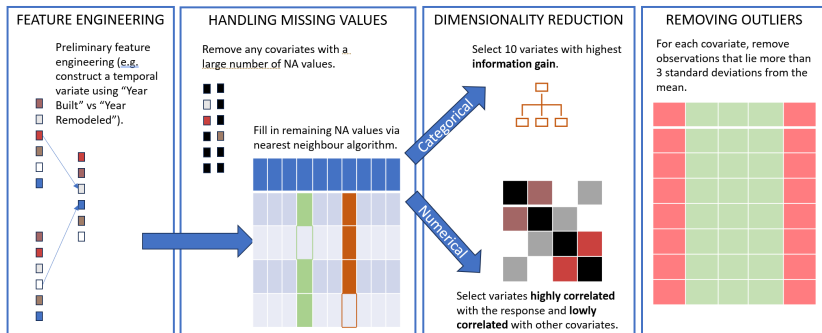# STAT 444 Final Project: Ames Housing Dataset Regression Analysis

Angelo Carreon, Hoseok Lee, Joy Chen, Steven Shen

Exploratory Data Analysis and Preprocessing

Metadata:

- In the original dataset, 82 covariates, $n = 2930$ observations
  - 23 nominal, 23 ordinal, 14 discrete, 20 continuous
- After data pre-processing, 19 covariates.
  - 7 nominal, 4 ordinal, 4 discrete, 4 continuous
- Response 'SalePrice' was log-transformed for a more normal distribution.

| FEATURE ENGINEERING	HANDLING MISSING VALUES	DIMENSIONALITY REDUCTION	REMOVING OUTLIERS
Preliminary feature engineering (e.g. construct a temporal variate using "Year Built" vs "Year Remodeled").	Remove any covariates with a large number of NA values. Fill in remaining NA values via nearest neighbour algorithm.	Select 10 variates with highest **information gain**. *Categorical*. *Numerical*. Select variates **highly correlated** with the response and **lowly correlated** with other covariates.	For each covariate, remove observations that lie more than 3 standard deviations from the mean.

Methods & Models

We compared four regression techniques to identify the most suitable model:

1. Multiple Linear Regression
2. Ridge
3. Lasso
4. GAM

We assessed their performance using a 20-fold CV scheme, then tested their final performance under an 80:20 Train/Test split.
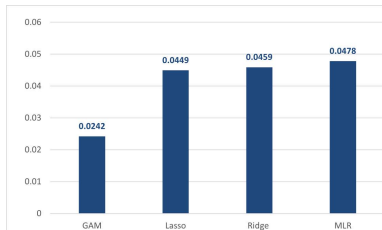
(a) CV Score
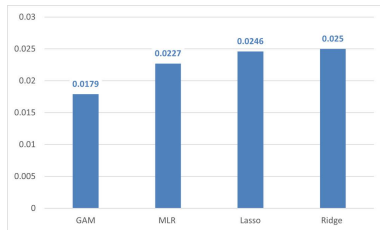
(b) Test Score

Figure 1: Comparison of Model Performance

Final Model Analysis: GAM Model

The GAM model appeared to generate the best predictions. Why?

- Lots of non-linearity in the final list of (continuous) covariates
- If it was addressed, we would expect better regression performance
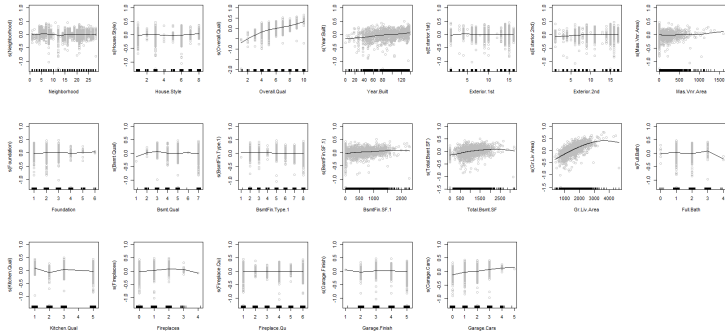- In the end, our predictions were accurate with RMSE of approx. $18,000

Figure 2: Fitted Smooths and Plotted Data of the final GAM model