### Team members

* Mriganka Sarma (netID: ms76) – <<add some contributions info here>>
* Ajay Menon (netID: kamenon2) – <<add some contributions info here>>
* Kai Pak (netID: kaipak2) – <<add some contributions info here>>

### Ames Housing Dataset - Exploration / Analysis

## Objective

The objective of this project is to evaluate different models with the given Ames dataset and then measure the accuracy for the generated models. In order to calculate the model accuracy we use a general-purpose error metric called RMSE (root mean squared error) which helps us calculate the prediction errors of different models. We select test ID’s from the given file (project1\_testIDs.dat) and on each of 10 iteration’s we calculate the error and save the results in the end.

The given original data set contain 2930 observations overall and 83 columns. We have also been given the test ID’s set which help in construct for the train and test dataset accordingly. Initially we started with our basic analysis of the data set and generated plots to understand the datasets better. This gave us an understanding as to how the data preprocessing would need to take place. Note: Since RMSE is good with numerical predictions and since we have categorical data included in the data set, it is required that we preprocess and reengineer the variable’s accordingly.

## Basic Statistics

The tabular format gives an overview of the dataset and the missing columns and observations for one of the test ID series (j = 2) and helped us understand the data sets better. We see the missing data observations and the outliers which create models to misbehave.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Full Data Set | Train Set | Test Set |
| Rows | 2930 | 2051 | 879 |
| Columns | 83 | 83 | 82 |
| Discrete columns | 46 | 46 | 46 |
| Continuous Columns | 37 | 37 | 36 |
| All missing columns | 0 | 0 | 0 |
| Missing observations | 159 (0.065%) | 0 | 0 |
| Complete Rows | 2,771 (94.53%) | 2051 | 879 |
| Total observations | 243,190 | 170,233 | 72,078 |

Chart, bar chart

Description automatically generated

Outlier for the Garage Built Year’s need to be handled.

## Data Pre-Processing

1. Preprocessing: Test and Train CSV creation
2. Variable Removal:
3. Imputations: Missing data profile – only garage belt seems to missing records out of the 83 variables and seems to be at 5.43% of its number of observations.
4. Dummy Variables (dummyVars – Caret Package)
5. Winsorization
6. Log transformation (Sale Price)

## Fitting Prediction Models

##### Preliminary Analysis:

* <<Started with Lasso and Random Forest Implementations as the 2 models which needed iterations>>
* <<RMSE findings for 10 iterations and calculated the train and test set thresholds>>
* << Variable adjustments >>
* << 2 by 10 accuracy table here>>
* << Running Time – Total time for training and test on two models for each iteration>>

##### Final Implementation Details:

* <<Elastic Net (Linear Regression Model) + XgBoost (Tree-based Model) Implementation Details>> or <<Lasso+Ridge and XgBoost Implementation Details.>>

## Discussion’s

<<Interesting finds if any>>