Final Project Write-Up

Group G: Sara Culhane and Brenna Sullivan

Data:

For our final project, we used the dataset from the Kaggle Competition House Prices: Advanced Regression Techniques to try to fit a model that would best predict the sale price of houses in Ames, Iowa from 79 predictor variables that assessed the location, quality, square footage, building material, and several other factors of the houses. The data was split by Kaggle into a training and test, and we further split the full training data into a test and a train for our pre-submission testing. Figure 1 shows the spread the log of sale price of the houses.

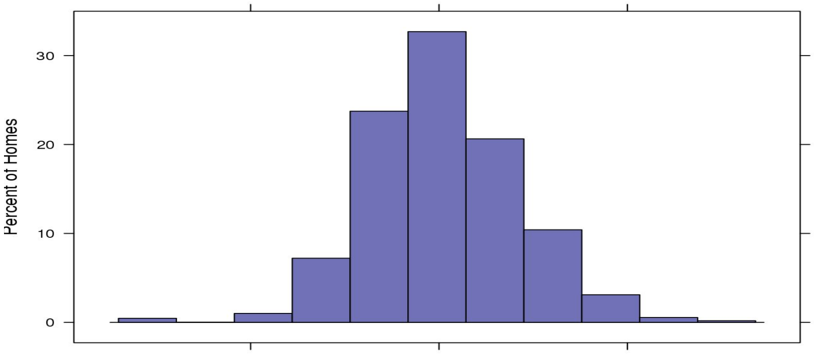


Figure 1

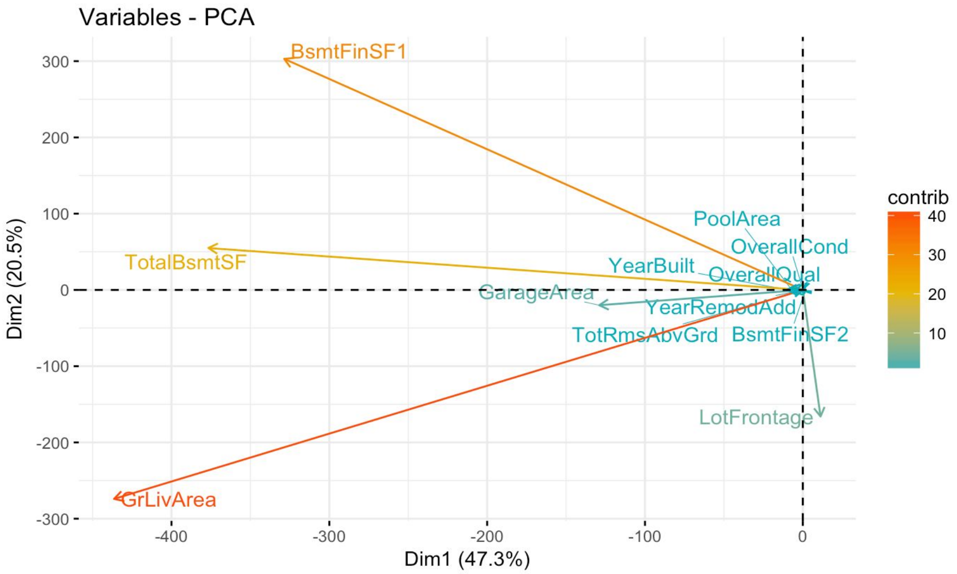
Initial Methods:

The first method we used for predicting the sale price of houses was stepwise regression, which gave us an output that created a linear model using nearly 40 predictor variables. While effective on our subset of the test (RMSE of 0.09), we could not get an accurate read from our kaggle_test dataset because there were too many NAs in the set. Despite efforts to wrangle the data, the NAs greatly impacted the model, so we moved on to other methods, while keeping the variables from the stepwise output in mind.

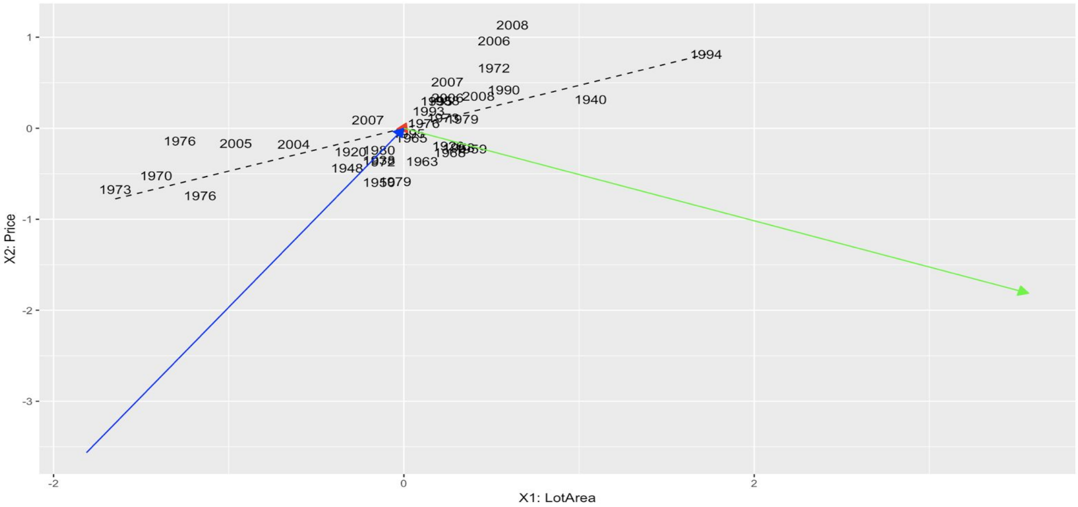
Another method we used to fit a model to predict the sale price of houses was LASSO. LASSO performed very well on our test set split from our training with an RMSE of 0.08, but we were unable to reproduce this on the full set due to similar issues with missing values that we faced with stepwise regression. When we tried to remove problematic columns as we discussed in office hours, we still faced issues with the dimensionality of the predictor matrix, which only retained 1339 of the 1459 rows when dropping these columns. Had this worked on the the full set, it likely would have been our best model in terms of predictability.

Final Submission:

The method we implemented and used for our final submission was Principal Component Analysis, which helped us to see the how well some predictors explained variation and to see "interesting transformations" through our plots. Particularly, we found that a set predictors created the following set of principle components:



We also took a look into price and lot area with year built (below) as a second variable to test which of these predictors explained the most variability. Ultimately, we left both out and created a linear model using the variables observed in the preceding plot.



Though we ran also into issues with the kaggle_test set once more, we replaced the missing values with the mean value and received a fairly respectable RMSE from Kaggle of 0.22, which was about twice the size of the RMSE found for our training data to be.

If we were to do this over again:

We would likely pick a dataset/kaggle competition with a more robust test set or perhaps would have chosen a method that was better with dealing with these NA's. PCA did an okay job with this, but it seems that the missing values impacted the model we generated with its features as well. LASSO and stepwise regression both struggled to deal with this issue. If Kaggle allowed for the submissions with fewer dimensions, that might also impact our model's effectiveness.