Ben Buzzee, Thanh Nguyen, Kellie Mcclernon

Team name: NotTooDeep\_Learning

Weekly Summary 1

April 7, 2017

This week our team focused on exploring and cleaning the Ames Housing data. We found that most of the NAs present had known values mentioned in the data description document provided by Kaggle. After replacing the NAs with known values, we still had a handful of missing values to deal with. To allow us to fit models we replaced those missing values with medians for quantitative variables or modes for categorical. More work will be needed to ensure replaced missing values are well chosen.

After fitting a handful of individual models, we decided to submit our best individual model to provide us with a baseline RMSE we can strive to improve. The generalized boosted regression model ('gbm' in caret) provided the best repeated cross-validation RMSE at .1307 (for log transformed saleprice). Our submission results were surprisingly similar with a RMSE of .13066. Our current rank is 1161 out of 2245 teams. Now that we have a "clean" dataset, we can now focus on dimension reduction, feature engineering, and combining models.