Joshua Kristanto

Jose De La Rosa

Albert Eng

Feature-Engineering -2

Goal: Improve previous scores of 0.23, 0.44 and 0.20 using feature engineering.

In our feature engineering we used all numerical type columns. A few columns that were not supposed to be numerical passed through so instead of dropping we decided to 1 hot encode them. This column were MSSubClass. I also one hot encoded whether the house was remodeled at some point in its lifetime, assigning a 0 if it was not remodeled and a 1 if it was. Another column that we engineered was if it was old, I used the definitions of Kaggle, found under MSSubClass column description.txt, where it considered greater than 1946 new and anything below that as old. Using that information, I set up a new column telling us whether the building was new or older. MSSubClass was 1-hot encoded by the building type, I used 1 and 1-1/2 as a category, 2 and 2-1/1 as another, split and multilevel as another and used duplex and multilevel as the fourth different, after assigning these tags to the respective building I 1-hot encoded them. I also created an interval scale for how old the house is, I used 2010 as it was the year of the posting. I added the several porch square foot columns to make a total porch sf variable. I did this to both Basement baths and Baths and their relative to condense them into two different variables, “Baths” and “BsmtBath”.

With these features we used a Random Forrest Classifier, Random Forrest Regression, Multilinear Regression. In the end when we submitted to Kaggle our best score of 0.18 using the Random Forest Classifier, our Random Forrest Regression got a score of 0.21 and the Multilinear Regression got a score of 0.71.