IA1 Kaggle Competition Feature Engineering Report

I explored the following feature engineering techniques:

1. First, I converted the categorical variable “zip-code” to multiple distinct dummy features each representing different distinct zip codes. By hot encoding the zip code feature only, I noticed an improvement in the MSE of the training set. I also dropped the “id”, “sqft_living15” and “zipcode” features subsequently as they are redundant features adding no new information to the dataset.
2. Secondly, I explored creating several additional features out of the existing features. These includes making a feature column to represent unused residential space, calculated as the difference between the living space and the entire lot size. This simply represents how big of space individual residences have. I also created a feature to represent how long the residence have been built, to represent the age of each house. The significance of these new features are pretty intuitive since those are key factors that are taken into consideration when buying a new home. Some of the other new features created include a feature to represent the geographical location of individual apartments as well as a feature to denote if a residence is the largest in its neighborhood. When these new features are combined with the other feature engineering done from part 1 of IA1, I noticed significant improvement in the MSE of the training set.
3. The third technique that I explored did not yield an improvement on the first two. This technique involved using only the positive weighted features from the initial training using the batch gradient algorithm. This is done on the assumption that positively weighted features will have better correlation with price. However, I noticed the MSE increased very slightly showing that using only a subset of the features did not lead to any improvement on our training dataset.
4. Other techniques I tried included auto encoding the “grade” feature, but this did not lead to an improvement in the training MSE. Instead, the MSE increased. Similarly, I added up “sqft_above” and “sqft_basement” features to create a new feature, since they both indicates the interior housing space. However, it did little to improve the MSE and so I left it out as a feature.