Lab 7 Problem 3: XGBoosting

We scored a RMSE of .1247 in the Kaggle Competition. XGBoost did not end up getting us a better score than our Linear Regression with Ensemble Lasso and Ridge Estimators. I found a useful document that helped me implement and understand XGBoost in python. It was helpful in seeing their use of one hot encoding for their variables and being able to relate that to how I used the pandas function of get dummies. There was also another helpful code that showed the use of removing skewness from our data. We could see from our results that some features were very skewed and we were able to counteract this by taking the log of these skewed features. Using XGBoost, we were alone unable to get a better RMSE. However, combined with K-fold cross validation, I was able to see large improvements in our score. We also found that higher k-values resulted in a better score. A k value of 100 was significantly better than that of 10. I tried treating the problem as a classification with many leaf nodes to separate different price ranges through label encoding. This worked decently well but was not as good as the Regression Approach.