Title: House Prices – Advanced regression techniques
Group name: ML group
Current rank: 340
Group members: Gnaneshwar Mandava(df5864), Henry Lao (gm4427)

Submission 1: Rank 6000

First skewness is calculated and replaced all the null values with "None" and changes the categorical values to numerical. As there are numerous features about the garage and basement, fill some with "None" and rest with the mode of that feature. Combining the square foot columns into the "totalsquarefoot" and Calculating age using the year of sold and built. Now, dropping the combined columns, and filling the rest of the data with the dummies and the skewness is calculated. For the highly skewed features, log transformation is applied, and the data is now divided into the train and test variables. In which GridsearchCV is used to find the best parameters. Once the best parameters are found, those parameters are applied for the XGBoost model regressor. Then the calculated RMSE value is 9.1096 which was not the best result.

Submission 2: Rank 1600

No changes were made to the pre-processed data but GrissearchCV was used again to find the find the parameters for the Support vector machine (SVM). Now the model is applied with the calculated parameters and then the result is obtained. The RMSE calculated as 0.1124

Submission 3: Rank 1600

No changes were made to the pre-processed data but, for this submission I've used the deep-learning techniques with the TensorFlow, with weight normalization and using model checkpoints with Batch normalization and with the best weights are added and the predictions are done. The rmse is calculated and it is 0.1236

Submission 4: Rank 340

No changes were made to the data preprocessing but the stacked models are added which contains the Xgboost, LightGBM and Elastinet are added and then and the predications are calculated and then the RMSE is 0.11945

Submission 5 : rank 340

The first submission of the XGBoost was a model using parameters found using a grid search with a score of 0.14355. Any missing values in the training and testing data sets were filled in using the data_description.txt. XGBoost was the first model used as it was related to the previous assignment as a result of some familiarity from the most recent homework leading up to the kaggle project. Furthermore, XGBoost appeared to be one of the few popular models used in the

competition and peaked interest in benchmarking the model. No feature engineering was applied.

Submission 6 : rank 340

The sixth submission performed poorly therefore drove interest into further researching the methods applied by fellow kagglers. Sifting through the various discussions and uploaded kernels, various models were found such as Lasso, Ridge, Support Vector, GradientBoost, XGBoost and LightGBM to be applied in an ensemble model. Furthermore, fellow kagglers were applying techniques such as feature engineering in creating new features as well as what appeared to be better imputation techniques; thus the goal was to attempt to create an ensemble model using Lasso, Ridge, Support Vector, GradientBoost, XGBoost and LightGBM algorithms. Feature imputation was revised. Specific numeric features were normalized using a log transformation as well as squared. The resulting score was 0.12244.

Submission 7 : rank 340

The seventh submission was significantly better than the first submission when using an ensemble stacking model. Researching other notebooks to further improve the score, I found that a few people were applying final adjustments to predictions of a particular quantile. The only adjustment made to this submission was an application of the formulas utilized by other notebooks where the 99% and the 1% percentile were adjusted. The result was a marginally better score of 0.12240.

Submission 8 : rank 340

The eighth submission was driven by explorative reasons and a motivation to simplify the ensemble and to use fewer models. The resulting ensemble consisted of a stack regressor as the base model mixed in at 0.678. The meta models consisted of a Ridge and a XGBoost model both mixed in at 0.1665. The resulting score was 0.12428 marginally worse than the 3rd iteration.

Submission 9 : rank 340

In the 9th submission , the extra tress regressor is used and then along with the stacked regressor and which contains the Xgboost , Gradient Boosting and Light gbm with the Lasso as the base and then the calculated RMSE is 0.1236

Submission 10: Rank 340

In the 10th submission the same model as the 2nd submission is used but with the RidgeCV but there has been no improvement and have calculate an RMSE value of 1228.

Submission 11: Rank 335

The intuition leading to this score was tuning of the blend of the ensemble, xgboost and lightgbm. The starting ensemble model was made up of a base consisting of ENet, GBoost,

KRR, Ridge and a meta defined by Lasso. The blended model consisted of 0.7 Stacked Regressor, 0.15 XGBoost, 0.15 LightGBM with a score of 0.11961.

Submission 12: Rank 322

Motivation here was to adjust the blend. Here improvements were noticed after decreasing the starting blended model to 0.7 Stacked Regressor, 0.10 XGBoost, 0.15 LightGBM with a score of 0.11941. This implied that XGBoost was not contributing negatively to the overall performance.

Submission 13: Rank 317

The next intuition behind the this step was to completely remove XGBoost such that the final blended model consisted of 0.7 Stacked Regressor, 0.0 XGBoost, 0.3 LightGBM. Increasing the dominance of LightGBM in the meta was motivated by controlling the Stacked Regressors' contributions to the overall model. At this point, there are only 2 models to blend: Stacked Regressors and LightGBM. The following score of 0.11904 was achieved by adjustments to the Stacked Regressors' contributions where the Stacked Regressors contributed to 0.65 and LightGBM contributed 0.35.

Submission 14: Rank 317

Reasoning behind this step was to confirm exhaustion of maximizing the trained models. This was confirmed by testing a 0.55-0.45 blended model yielding a score of 0.11905.