Foundation Of Data Science Project Report: Ames House Price Estimation

1860273 Joodi Bigdello Manoochehr
1869394 Majidi Emad

Team name:
MJ.1860273_EM.1869394_FDS
Score: 0.11619

- **Language: Python**

- **Data tidying:**
    1. Removed rows with > 60% Nan (Alley, PoolQC, Fence, MiscFeature)
    2. For categorical put None Instead of Nan's, and for numeral's put mode, median or '0' based on the values.
    3. Removed outliers Based on scatter plot of every 'feature value' and 'target value'.
    4. Normalization: for features with skewness >60%, we do log + 1 transform.

- **Feature engineering:**

1. Delete Features with correlation value > 80%. (GarageArea, 1stFlrSF, TotRmsAbvGrd, GarageYrBlt, KitchenAbvGr)
2. Create 'Dummy Variables' for categorical ones.
3. Add Poly (poly = 2) Feature for 12 most important Features. (Based on lasso)
4. Add feature 'NewHouse' (1 for new house and 0 for others) based on 'YrSold' and 'YearBuilt' features subtraction.
5. Add feature 'OverallSF' based on sum of '2ndFlrSF' and 'TotalBsmtSF' features.

- **Base Models:**
    1. One LassoCV
    2. One LassoLarsCV
    3. One Elastic Net
    4. One Linear Regression
    5. Two Random Forest Regressor
    6. Two Gradient Boosting Regressor

- **Training**
    1. Run every base model with 10-fold KFOLD validation and do prediction for test data.
    2. Then use average of base modes to create the model.
    3. After creating final model we use GridSearchCV from sklearn as stacking model with Ridge as estimator and find best hyper parameter for this estimator.
    4. We do prediction based on GridSearchCV and achieve 0.1012 in python test and 0.11619 in Kaggle website.