FDS Fall 2017 - Kaggle Project

Student : 1388371
Name: Umberto Junior
Surname: Mele

Kaggle: https://www.kaggle.com/drago91
Kaggle account: UmbJrMl

Data tidying:

1. removed outliers with 'GrLivArea' > 4000 & ''SalePrice' < 300000.
2. Normalized 'SalePrice' using log plus 1.
3. Clustered variable 'Neighborhood' in four class.
4. replaced some outliers with Mode or other.
5. Rounded at basis 10 or 100 float features : 'BsmtFinSF1', 'LowQualFinSF' and 'WoodDeckSF'.
6. replace dna with 0 , ''None' or mode.
7. Substituted 'GarageQual' with ordered numbers.
8. Transformed 'MSSubClass', 'OverallCond', 'YrSold' and 'MoSold' to string
9. Label Encoded some features using CV on differents models
10. Standardizing all skewed numerical features with box-cox 1 plus
11. used RobustScaler in the Pipeline

Feature engineering:

12. creation of ''TotalSF' total square footage feature

Feature selection

13. removed 'Fence' and 'Utilities' columns since the regression there was a better score during the CV
14. two differents train data has been build one for lasso based models and for boost based models both selecting differents features

Model Selection

15. Using CV 8-fold on many models: 'Lasso', 'ElasticNet','Ridge', 'BayesianRidge', 'HuberRegresion', 'XGBoost','Lightgbm', 'LinearGAM', ecc.. best scores are:
    a. Lasso score : 0.1089 (0.0121)
    b. ElasticNet score: 0.1087 (0.0122)
    c. HuberRegressor score: 0.0936 (0.0142)
    d. GBoost score: 0.0980 (0.0147)
    e. Model XGB score: 0.0987 (0.0122)

16. at least some model are been averaged to create a new model, and using stacking my best prediction is:
    a. Is a average model of ElasticNet, GBoost and HuberRegressor
        i. Kaggle score:0.11510