1831467 Almagout Nagham team name: Theta
1735885 Ismail Hassan score: 0.11335

Language: Python

Data tidying

1. We didn’t removes any rows with NA values instead of that we replaced NA values with 0 for a group of features, with ‘missing’ for other features, and with mode values for the rest of the features, in addition some of them we fill NA values by predicting its nan values.
2. We removed 3 rows with 3 outliers, during our tries we removed 4 and 6 rows instead of 3, but we found is better to remove just three that relate to the outlies of the columns “OverallQual” and “GrLivArea”
3. We normalize the numeric data using “log1p” function.
4. We convert categories data to numeric data.

Feature engineering

1. We created some new features such as:
Total Place which is the total square feets
Porch Qual which is OpenPorchSF*OverallQual, TotArea shich is '1stFlrSF'+ '2ndFlrSF'+'TotalBsmtSF',
BsmtFinSF1Qual which is 'BsmtFinSF1'* 'OverallQual',
Bsmt which is "BsmtFinSF1" + "BsmtUnfSF")* "OverallQual",
Rooms which is "FullBath"+”TotRmsAbvGrd"* "OverallQual",
LotArea which is the LotArea*OverallQual,
and other features like HasPool if the house’s pool area > 0.
2. We drop some features we noticed that it not effect on the interested feature and contain a lot of NaN values such as: 'MiscFeature', 'Alley', 'Fence'. In addition to some features that we used it in creating new features like: '1stFlrSF', '2ndFlrSF', 'TotalBsmtSF', 'BsmtFinSF2'.
3. We tried to see if multicollinearity problem affect our model, where we have in our dataset many variables that correlate to each other, so we tried to remove one variable from each pairwise correlated variables but that did not improve our score.

Medialization

1. Lasso
2. Gradient Boosting Regressor
3. Linear Regression
4. Ridge
5. Kernel Ridge
6. ElasticNet
7. LGBMRegressor
8. XGBRegressor

Training

1. Split data on 30-folds cross validation and repeated six times, we tried different splits but this was the best one.
2. We get the mean of our models predations with different datasets each one with different number of features depending on Lasso model selection.