House-Prices : Advanced Regression Techniques

Sold!! How do home features add up to its price tag??#KAGGLE COMPETITION

Objective:

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home. The potential for creative feature engineering provides a rich opportunity for fun and learning using basic and combinations of regression techniques.

Data Exploration :

- The train and the test data is concatenated.
- NaN values are replaced with significant values.
- The categorical variables which have some order(e.g. Poor, good, excellent) are replaced by numeric values [0, 1, 2, ...]
- Yes/No in certain features are replaced by 0, 1.

New Feature Creation:

- The feature extraction was quite minimal taking log transformations of numeric features and replacing missing values with the mean.
- It Didn't worked well - To improve it we added New features to it.
- Ordered categorical featured are divided into good and poor. The idea is good quality should rise price, poor quality - reduce price.
- Exterior1st, Exterior2nd, RoofMatl, Condition1, Condition2, BldgType are converted to price brackets using SVM.
- All the features are added to the original feature set, combining with different combinations of original features.
- The final shape of the training data set was 1459 * 460 columns.

Scaling Features

- First we have transformed the skewed numeric features(skew value is greater than 0.75) by taking log(feature + 1) - this will make the features more normal.
- Logarithm is applied to target value as well.
- Dummy variables for the categorical features are created.

- The insignificant features are removed, which have many zeroes.
- Missing values are replaced by mean whereever left.
- The outliers id's are identified and dropped.

Models

We have used 3 different types of models :- - #Linear regression with Lasso Regularization - The idea is to try Lasso a few times on bootstrapped samples and see how stable the feature selection is. As we have mostly tuned the parameters and Lasso should perform better. So giving more weight to it . - #Xgboost model to our linear model to see if we can improve the score giving weight to its prediction - #Elastic net Which is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods.

Prediction

- No single model gave efficient outputs.
- Have tried with combinations of different models using basic ensemble methods.
- Final prediction is done with giving different weights to the three models as: w1*lasso_pred + w2*elastic_net + w3*xgboost where, w1+w2+w3 = 1

Libraries and Packages Required :

- Skitlearn `>= 0.18`
- Numpy `>=1.11.2`
- Pands `>=0.18.1`
- scipy `>=0.18.1`
- Xgboost `>=0.6`
- seaborn `>=0.7.1`
- More imports are :- itertools , operator