DataWarehousing and Data Mining Report
Team: Kaggle_

Indranil Mukherjee - 201505531
Munmun Chowdhury – 201506593

Solution Approach

Data Preprocessing:

1) Concatenate Test and Training set so that all the preprocessing steps will be consistent

2) Handling Missing Values:

For some columns NaN were replaced with most common and some with median and some with mean.
We checked different combinations and picked the one that gave the most optimal result.
For some columns like 'GarageBuilt' replacing NaN with the above options made no sense so we replaced them with 'No Garage'.

3)Normalization:

Handling Categorical Data-

Some features which were categorical but they represented some kind of order and thus could have been treated as numerical data.
Like 'FirePlaceQu' had values like 'Ex','Gd','TA','Fa','Po' and 'No FirePlace' which represented quality as excellent, good, poor etc.
Such values with replaced with scores like 0,1,2,3,4,5.

Similarly some features like 'Overall Quality', 'BaseMentQuality' has an effect on price as high quality tends to increase price whereas poor quality tend to decrease price.
Hence these features were divided in good and poor.

Some seasons(Apr-Jul) have high sales while some had low.
So these months have been taken into consideration.

Nomalized the Year field

Scaled Numeric data.

Dropped a few columns based on the criteria:

Columns that had very few unique elements
Columns that had almost all zero elements

4) Outliers:

Removed a few prominent outliers

Learning Models:

For this project we tested a few learning models, namely
i) Linear Regression
ii) Ridge Regression
iii) Lasso Regression
iv) XGBRegressor

Combining Lasso and XGB Regressor as
Score = 0.7*(Lasso Score) + 0.3*(XGB Score)
we got an rmse value of 0.11595 on the leaderboard

Combining Ridge Regressor to the above as

Score=0.6*Lasso + 0.3*Xgb + 0.1*Ridge
we got an improved score of 0.11568

We tried ElasticNet model along with Lasso and xgb
we got further improved score of 0.11546

changing the parameters in xgbRegressor we got the final score of 0.11528 on the leaderboard.

Best Rank obtained: 14