

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 were 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 tends 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.

Normalized 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

$$\text{Score} = 0.7 * (\text{Lasso Score}) + 0.3 * (\text{XGB Score})$$

we got an rmse value of 0.11595 on the leaderboard

Combining Ridge Regressor to the above as

$$\text{Score} = 0.6 * \text{Lasso} + 0.3 * \text{Xgb} + 0.1 * \text{Ridge}$$

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

#	Δ1w	Team Name	Score	Entries	Last Submission UTC (Best - Last Submission)
1	—	ThúlioCosta *	0.07832	24	Tue, 06 Sep 2016 01:47:15
2	new	HungryFools	0.10051	10	Thu, 06 Oct 2016 07:41:20 (-0.1h)
3	↓1	DoingItWithModels	0.10345	30	Tue, 20 Sep 2016 03:35:27
4	↓1	andrewwoohoo	0.10683	3	Mon, 12 Sep 2016 19:23:52
5	new	Attempts 2	0.11436	10	Thu, 06 Oct 2016 15:41:28
6	↑3	Justfor	0.11461	31	Mon, 03 Oct 2016 06:44:20 (-31.7h)
7	new	Apocalypse	0.11461	15	Thu, 06 Oct 2016 15:56:52
8	new	Wingardium Leviosa	0.11475	19	Thu, 06 Oct 2016 13:24:46 (-1.4h)
9	↑7	persistence	0.11490	22	Thu, 06 Oct 2016 13:58:35 (-13.4h)
10	new	Raj Patel 2	0.11498	10	Thu, 06 Oct 2016 09:30:44 (-0.1h)
11	new	emohinstead	0.11511	8	Thu, 06 Oct 2016 08:50:40 (-1.6h)
12	new	Hitesh Sharma	0.11511	13	Thu, 06 Oct 2016 08:56:51
13	↑80	shining_outliers	0.11523	10	Thu, 06 Oct 2016 16:03:41 (-3.2h)
14	↑64	Kaggle_	0.11528	24	Thu, 06 Oct 2016 18:12:02

Your Best Entry ↑
You improved on your best score by 0.00018