Fundamentals of Data Science - Final Project

Andrea Formichetti ID-1764939

09 February 2017

In this challenge we are being asked to produce a prediction about sale prices of residential homes in Ames, Iowa. Train set and test set needed for the prediction is provided as two separated CSV file, we want to produce another CSV file containing for each row the Id of the house and a prediction for the sale price. Follow the procedure adopted.

1 Preprocessing

We wipe the dataset as much as possibile in order to make our predictions more accurate.

Make trainset bigger In order to get better accuracy on the final prediction, I cut out the price column from the trainset and merge the two datasets together this result in a larger dataset and hopefully better estimations.

Dropping features with too many NaN I drop all features that have more than 40%(empirically chosen) of NaN among its entries. This features are too much uninformative and it would ruin the prediction.

Remove outliers In this phase I remove from the trainset all entries identified as outliers, in particular all those with a value that exceed the standard deviation more than 5 times. That's because linear model regression are quite sensible to the outliers.

Compute the skewness and take the log I detect the features that has an high skewness and i transform by taking $log(feature + 1)$ in some sense it make the feature more "Normal"

Generate dummy variable A lot of features among the 79 that compose the dataset are categorical, which is it take value from a finite list of string representing some quality, we need to convert it to something that can be inserted in our model. Generate boolean variable for each possible value is a way and in this case work well.

Filling NaN with the mean