Small report 2nd place Business Game 2018

Michele De Vita

April 17, 2018

EDA

Correlation between categorical variables

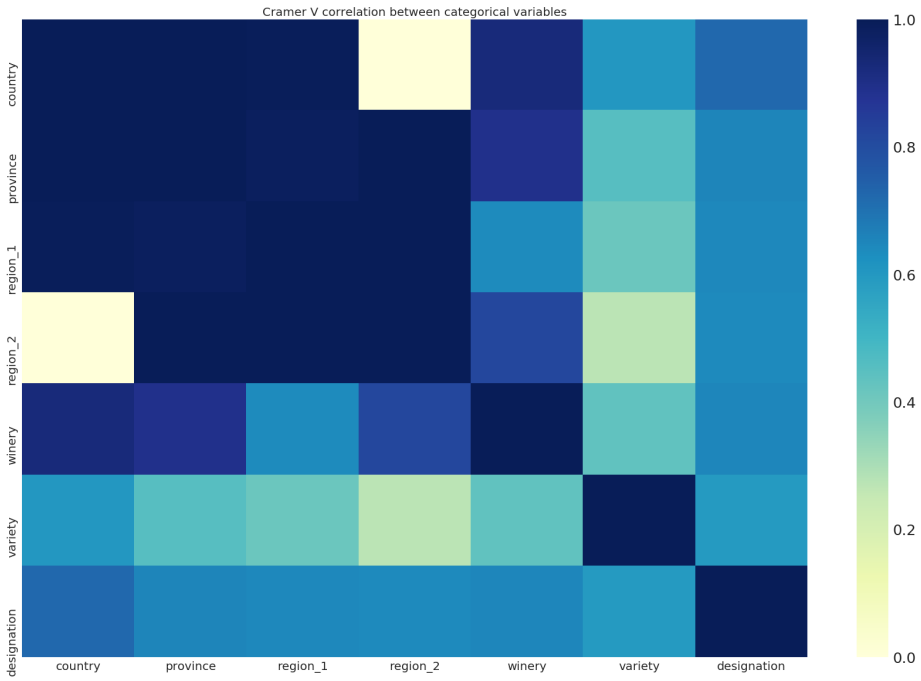

Figure 1: Cramer V correlation between categorical variables

Obliviously there is a strong correlation between *country, region, province* because they are all related to location where the wine is produced. There are few variables so i decided to not remove any correlated variable.

Preprocessing

I've tried many simple things but in 4 hours but the best i've used in my last submission is simply deleting two columns: *review* and *region2* because contains many *NA* and i've think that it can penalize my model (Random Forest). A

very interesting technique that i want to use for handling the text that i was unable to use is the word vector, in particular replace the text with the *norm 2* of the word vector of each text.

Model choose

For a 4 hour competition and a relatively small data my choose was the Random Forest because without too much tweaking on hyper parameters it learn very well from the data.

Validation of parameters of Random Forest

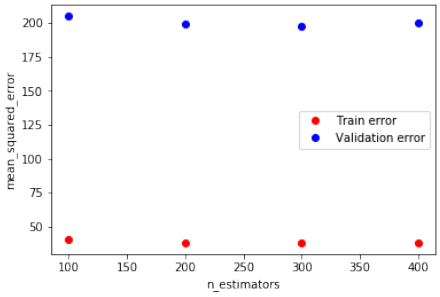

Figure 2: Validation on number of trees

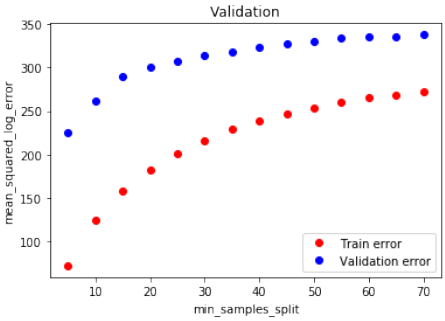

Figure 3: Validation on min sample split

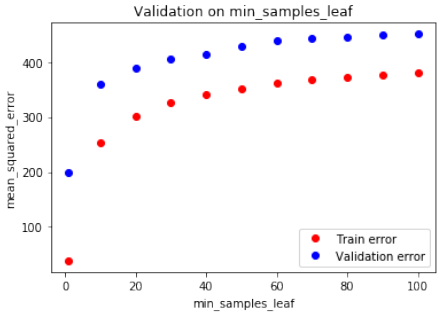

Figure 4: Validation on min sample leaf

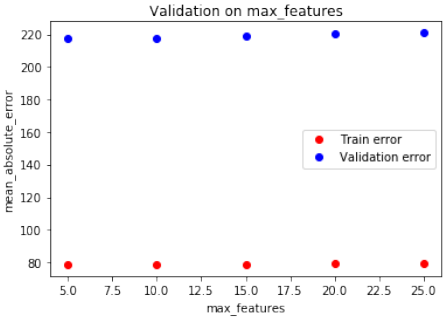

Figure 5: Validation on max features

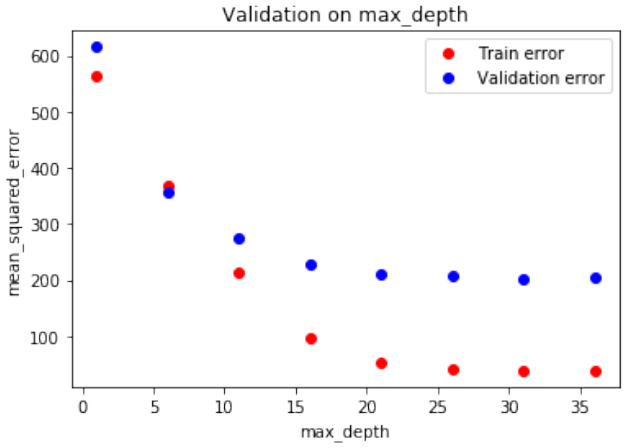

Figure 6: Validation on max depth

The only interesting parameter i found to improve is *max depth* that has brought an improvement on model prediction reducing the overfitting of the model. I

change from the default value 25 to 18 in order to improve the generalization of the random forest.