Wine Quality from content

René Angeles

March 16, 2020

> *He who knows nothing, loves nothing... But he who understands also loves, notices, sees ... The more knowledge is inherent in a thing, the greater the love. Anyone who imagines that all fruits ripen at the same time as the strawberries knows nothing about grapes.*
>
> —Paracelsus

1 Intro

We succinctly report findings on the wine quality dataset [1]. Our goals correspond to the data challenge launched at [2].

2 Exploratory data analysis

The EDA performed (see `EDA.ipynb` or `EDA.html`) shows that

1. Acidity predictors (fixed, volatile & citic), residual sugar, chlorides, sulfur dioxide (total & free), density , sulphates all have a heavy right tail. Naively, these tails are outliers. A log/log1p transformation gives distributions a bell shape, but in general the red/white wine distributions are shifted or different.

2. The correlations coefficients larger/smaller than 0.5-0.8. Hence the used of PLS methods is justified. Correlations coefficients are different between the respective white/red predictors.

Thus motivated, training was performed

1. With and without log transformed heavy tailed predictors. It was found that in all cases log transformation does not improves MAE.

2. With and without potential outliers where removed using a $\alpha \times IQR$ rule with $\alpha \in \{1.5, 1.8, 2, 2.2\}$. Yet, the removal of such potential outlier was never found to improve MAE.

No missing values/duplicates were found. TODO: Multivariate distributions could shed light relation between predictors. It is obvious that pH cannot be independent of acidity predictors, and similarly for other variables. Domain knowledge could actually help to tell how are these variables related.

3 Modeling

PLS is expected to perform better when predictors are center and scale Ref. [?, Kuhn, Kee] However, in our case we performed analysis without these transformation, finding that such transformation do not increase the MAE score.

Respectively, the files `Baseline.R, Pls_polr.R, Plsr_ensembles.R` and `GLMs_ensemble.R` are standalone models containing the feature engineering options presented above.

Main models reported		
Model.	Train MAE	Test MAE
Baseline: plsR, 5fold-tuned, independent models for red/white datasets. Predictions rounded.	0.4363	0.5026
pls-polr, 5fold-tuned separately for red/white sets.	0.5074	0.5181
Rounded prediction from a plsr stack of 3 plsr models trained into 4 separate train-sets (5fold tuned).	0.514522	0.5285054
Rounded PLSR stack of 3 PLSR models trained with resampling with replacement, 5fold tuned.	0.5074	0.5181
pls-polr blend of 3 models (pls-polr 5fold-tuned, pls-gaussian, psl). Only white wine was considered here	0.5080	0.5556

Note I did not included validation MAE. TODO: Figure out

4 Insights:

- The PSLR baseline was undefeated but it is worth noting the stacked ensemble has similar MAE for training and valiadation, i.e. it has lower variance.
- An important comment, due to the time constrain I didn't tune the hyperparameters for the models appearing in this ensemble, this is possibly why it performed poorly.

```
Coefficients:
                              [,1]
Intercept             3.896568e+01
fixed.acidity        -2.972064e-02
volatile.acidity     -1.871656e+00
citric.acid           5.856155e-03
residual.sugar        4.009412e-02
chlorides            -8.005123e-01
free.sulfur.dioxide   5.688777e-03
total.sulfur.dioxide -8.926095e-04
density              -3.757588e+01
pH                    3.125018e-01
sulphates             4.932548e-01
alcohol               3.280639e-01
Information criteria and Fit statistics:
                AIC    RSS_Y      R2_Y R2_residY RSS_residY
Nb_Comp_0 10184.357 3082.226        NA        NA   3918.000
Nb_Comp_1  9414.258 2531.059 0.1788210 0.1788210   3217.379
Nb_Comp_2  9012.642 2283.360 0.2591846 0.2591846   2902.515
Nb_Comp_3  8906.427 2221.173 0.2793608 0.2793608   2823.464
Nb_Comp_4  8896.766 2214.573 0.2815019 0.2815019   2815.076
```

Figure 1: Predictor importance of baseline

- When performing kfold tuning of the number of predictors to keep were lower 6. This means that at least 5 predictors suffer from collinearity problems.

5 Predictor importance

I detailed analysis of confidence intervals is `plsrglm` is possible but beyond the scope of the main concerns of the challenge perse. To don't go without intuition, we can can check at least the largest coefficient of the sucesfull baseline model, Fig. 1, which suggest that volatile acidity, density and clorides have the largest coefficient in the regresion.

6 Engineering

The code produced for this project has been refactored, and all the combinations tried above can be easily combined.

References

[1] Paulo Cortez *https://archive.ics.uci.edu/ml/datasets/Wine+Quality* .

[2] Analytic Flavour System *https://www.gastrograph.com/blogs/gastronexus/interviewing-data-science-interns.html* .

[3] M. Kuhn, K. Johnson *Applied Predictive Modeling* .

[4] Kee Siong Ng *A Simple Explanation of Partial Least Squares* .