Wine dataset Regression model

Nicola Scarano
Politecnico di Torino
Student id: s287908
s287908@studenti.polito.it

***Abstract*—In this report we introduce a possible approach to the *wine dataset* consisting into a heavy preprocessing on the categorical data. A deep analysis on geographical information and on ’winery’ was carried out to allow us to reduce the cardinality of these feature by continuing to retain important information. Also a word embedding approach has been used on ’designation’.**

I. Problem overview

We propose a solution for the regression problem of predicting the quality of a wine given some information about it. The dataset provided contain over 150,000 rows divided in development set and evaluation set. The first will be used for the training while the second is the test set where the final performances will be evaluated. The wine dataset is composed by eight categorical features plus the target class which has numerical values that range from zero to one hundred. Here we report all the attributes with a brief description:

- ’country’: nation of production of the wine.
- ’province’: That specifies a portion of a country i.e a state of the US, an Italian region or more general indication like: ’South Italy’, ’Northern Spain’ etc.
- ’region1’: refers to a more precise location, typically is a region inside the ’province’.
- ’region2’: is also a portion of a ’province’ but it refers to a region bigger than ’region1’.
- ’variety’: is the variety of the wine.
- ’winery’: is the denomination of the producer.
- ’designation’: is the denomination of the wine.
- ’description’: is a natural language attribute in which is reported a description/review of the wine.

Two main problem affect these features: they have missing value and they have a lot of unique values:

Feature	Missing values	Unique values
variety	0	603
winery	0	14105
country	3	49
province	3	456
region1	13889	1207
designation	25944	/
region2	50734	19
description	0	/

Because of the high number of unique values a one-hot encoding strategy with which to transform them is unfeasible except for a data transformation step. The outcome would be a dataset too big for the available hardware.

Four features out of the eight specify a geographical location, an information that we suppose related to the place where the wine has been produced or where the vineyard is located. An important concept is the size of the region considered both in terms of presence in our dataset in absolute terms.

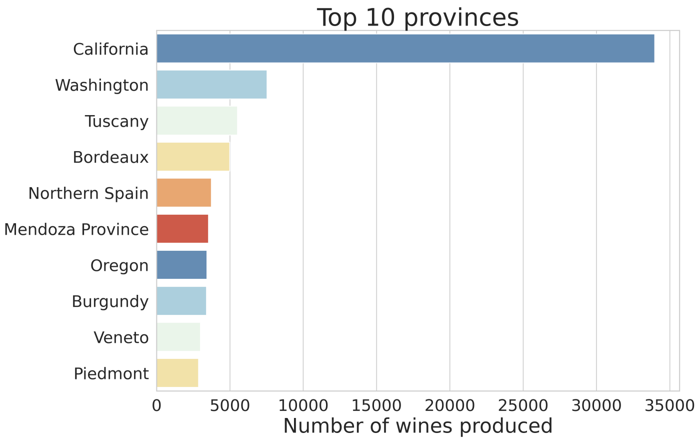


Fig. 1. Top 10 provinces for number of wines. We can see that in addition to producing many wines these province are also very extensive.

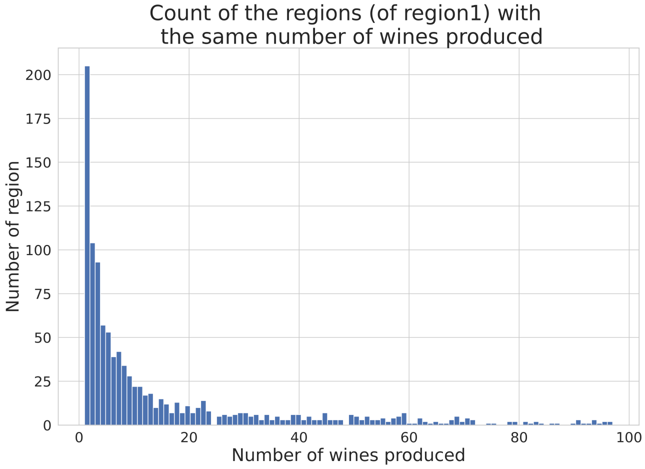


Fig. 2. In this picture we can see how there are a lot of regions in ’region1’ with few wines produced. Predicting the quality of wines based on this information would be wrong, as the information on which we would base the prediction is scarce and does not allow us to make correct generalization.

We can see that there are provinces (and also countries of course) to which many of the wines in the dataset belong. It is correct to suppose that this kind of information is useless for our task because it is too general. Too many wines of the most varied quality are produced in a large province or in a country. On the other hand geographical information more precise as the one provided by 'region1' help us to be more specific about the location but is likely to lose information. If a region is scarcely present in our dataset a prediction based on the quality of the few known wines of that region is likely to be affected by noise.

Regarding 'winery', its importance in the prediction of quality is certainly not to be doubted but the problem remains its more than 14000 distinct values that do not allow one hot encoding. The adopted solution here was to extract useful information hidden in the attribute.

II. Proposed approach

A. Preprocessing

Our preprocessing phase is organized in more steps:

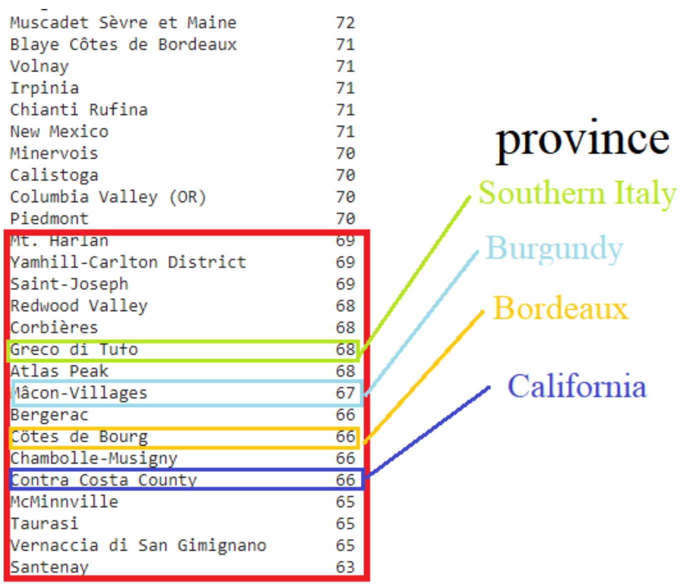
- Preprocessing on geographical information
- Preprocessing on 'variety'.
- Feature extraction on 'winery'.
- Word embedding on 'designation'.

Before computing the first step we remove duplicates and some noise found in the quality field of the development set: we delete all the record with 'quality' equal to zero. To encode the geographical information we choose to use the one hot encoding on a transformed version of the field 'region1'. The preprocessing of the geographical information consist in:

1) Fill the null values of 'region1' with the corresponding province.
2) Select a threshold α, minimum number of wines that a region have to produce to be used in the new region1 field.
3) Replacing in 'region1' whit the name of their province all the regions that appear lesser than the α.
4) Repeat step 3 but replacing the region with the corresponding country.

The idea that guided us was that we have to select the correct size of the geographical data such that to provide certain and complete information. With the threshold strategy we can find the best form of the geographical data and we also decrease the number of unique values so that we can perform a one hot encoding.

Applying the same strategy for variety we introduce a new threshold β. We assign a special value to all the varieties that appear in the dataset lesser than the threshold in such a way as to reduce its cardinality.



```
Muscadet Sèvre et Maine          72
Blaye Côtes de Bordeaux          71
Volnay                           71
Irpinia                          71
Chianti Rufina                   71
New Mexico                       71
Minervois                        70
Calistoga                        70
Columbia Valley (OR)             70
Piedmont                         70
Mt. Harlan                       69
Yamhill-Carlton District         69
Saint-Joseph                     69
Redwood Valley                   68
Corbières                        68
Greco di Tufo                    68
Atlas Peak                       68
Mâcon-Villages                   67
Bergerac                         66
Côtes de Bourg                   66
Chambolle-Musigny                66
Contra Costa County              66
McMinnville                      65
Taurasi                          65
Vernaccia di San Gimignano       65
Santenay                         63
```

Fig. 3. Suppose that the threshold for the geographical data is seventy. The figure show a portion of the values of region1 ordered by their frequency and in red are highlighted the regions that does not satisfy the constraint. As shown in the figure these regions will be replaced in 'region1' by their province.

For 'winery' we decide to extract some information from it. Practically we create 3 new features:

- winery_count: Count of the number of wines per winery.
- var_x_win: Number of different varieties produced per winery.
- ratio_variety_x_winery:

$$\frac{wine\_var\_winery}{winery\_count}$$

  where *wine_var_winery* is the number of wine of the specified variety produced by that winery.

On 'designation' the preprocessing step consists in taking the word presence matrix of all the words appearing in the field and adding the first n most frequent words as columns to the dataset. In particular we have used a vectorizer with a lemmatizer to extract the word, we have removed the stop words and we have applied a binary weighting schema.

In our model the field 'description' has not been used. We have tried a preprocessing step on it using the TfIdfVectorizer class of sklearn with a lemmatizer, removing the stop words, trying different weighting schema and different combination of the parameters of the sklearn class but nothing of it has worked. So we leave the sentiment analysis on this field for future developments because a more deeply analysis has to be done on it.

B. Model selection

We have done a comparison on a simple version of our preprocessed dataset between a default Random Forest and a default SVM and the Random Forest have performed much better than SVM. So we choose the Random Forest as definitive model for our task.

C. Hyperparameters tuning

There are three main sets of hyperparameters to be tuned:

- α and β respectively threshold of the preprocessing step for region1 and for variety;
- n number of word of designation;
- Random Forest parameters.

Assuming that the hyperparameters are orthogonal we have individually select them. Due to the high number of features used, this was the only feasible solution for selecting the parameters. We have first performed a default Random Forest with the features generated by the preprocessing step applied on region1 with different values of α and than with the features generated by the one performed on variety varying β. The same process has been applied to the selection of the number of word from 'designation'. Concerning the Random Forest hyperparameters we select them by running the Random Forest on a simpler version of our transformed dataset with a reduced set of features i.e. high thresholds.

Model	**Parameter**	**values**
RF with 'region1'	α	$300\rightarrow 50, step50$ $50\rightarrow 20, step15$
RF with 'variety'	β	$120 \rightarrow 20, step20$ $20 \rightarrow 10, step5$
RF with 'designation'	*n* max_df [binary, use_idf, norm]	[30,60,120] [0.7,0.35,0.15,0.5] [[True, False, False] [None, False, None] [None, True, None]]
Random Forest	n_estimator max_depth minImpDec	[50, 100, 150, 200] [None, 5, 10, 30, 50, 100] [None,0.1, 0.2, 0.4, 0.7]

III. Results

The best configuration for random forest was found for n_estimator = 150, max_depth=None, minImpDec = None. Concerning the other hyperparameters, α has been set to 50 that means 436 unique value of region1 while β to 20 and so 194 values for variety. Is important to say that lower value of the two thresholds allow us to have a too low performance improvement which does not justify the use of many more features. For *n* the best performances are guarantee by *n* = 120 with a binary weighting schema. We reach the best performance using also a one hot encoding on the one hundred most frequent provinces. This configuration allow us to reach on the public set an r2 score of over 0.79.

IV. Discussion

We are sure that the obtained result could be improved. First of all for the model selection: we have tried to use only two models, a Random Forest and a SVM and maybe more complex model as Neural Networks could reach better results. Furthermore a more precise preprocessing phase could be performed. A deeper analysis of the transformations applied to the geographical data, that we believe full of meaning, should be performed. As indeed also the for the features extracted from winery and for the transformation of 'variety'. A lot of new features can be extracted from the categorical data both from the geographical ones and from winery and variety. Moreover a more deeper analysis on 'description' can be performed. We have done something but a lot more can be tried and maybe some useful information can be extracted.

References