Wine quality regression problem

Valerio Zingarelli
Politecnico di Torino
Student id: s281586
s281586@studenti.polito.it

***Abstract*—In this report we introduce a possible approach to a regression problem based on wine quality prediction. The proposed solution consists in a concatenation of one hot encoding for categorical features and Term frequency–inverse document frequency for dealing with wines' descriptions. The proposed pipeline provides good results and allows to outperform the provided baseline.**

I. Problem overview

The proposed competition is a regression problem on a wine quality dataset which is composed of 150.930 different entries. It's divided into:

- a development set: 120.744 istances with a quality label
- an evaluation set: composed of 30.186 entries.

Each istance of the dataset is characterized by 9 columns:

- country: wine's country of production
- description: a brief description of the wine
- designation
- province: wine's province of production
- region_1 and region_2: wine's regions of production
- variety: wine's variety
- winery
- quality: quality of the wine. In evaluation set, quality column is not present

The goal of the competition is to build a regression pipeline in order to correctly predict wines' quality. Analyzing development test is quite clear that our dataset is quite unbalanced: as it's possible to see in Fig.1, majority of wines has a quality score between 25 and 75. Something important to notice is the massive presence of NaN values in some columns of our dataset. As shown in Fig.2, region_1 and region_2 have many NaN values, but also country and province columns have 5 NaN values too.

II. Proposed approach

In this section, you will present your solution. Please fill in accordingly.

You can use citations as follows: [1] (you can add BibTeX citations in the *bibliography.bib* file).

A. Preprocessing

B. Model selection

C. Hyperparameters tuning

III. Results

Here you will present your results (models & configurations selected, performance achieved)

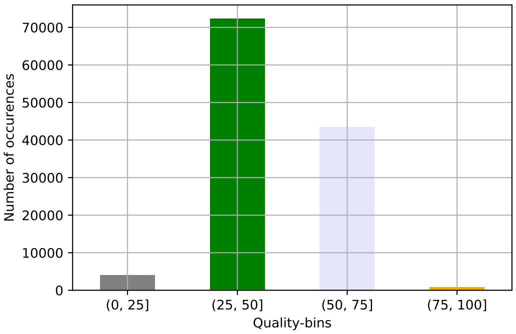

Fig. 1. Quality distribution

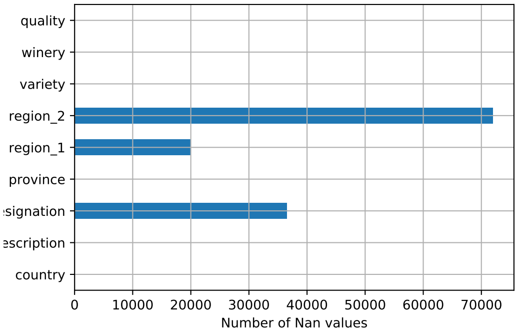

Fig. 2. Nan values distribution for each column

IV. Discussion

Any relevant discussion goes here.

References

[1] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.