Predicting Type and Quality of Wine From its Physical Conditions

Nerses Alikhanyan
Department of Applied Statistics and Data Science
Yerevan State University
nersoal@gmail.com

Abstract

Classification and rating of wines is interesting and time consuming task performed by highly trained experts. To move beyond our reliance on human experts several automation techniques can be applied once objective quantitative data of wine features is identified. This paper describes various machine learning approaches to solve the task of wine type and quality classification. Note that data of wine features used is obtained from objective physical measurements only.

1 Introduction

The global wine market is one of the biggest food industries worldwide. It is projected to grow with a CAGR of 5.8 %, during the forecast period of 2020 - 2025. Having such a huge presence in the industry wine factories still require significant amount of highly trained experts to estimate the quality of wines. Number of factors identify the final wine quality metric. Among such, apart from peculiar factors that could be qualified only by experts there are number of factors, estimation of which could be carried on through automation. In this paper we discuss such automation through well known machine learning classifiers. All considered methods use vinho verdo dataset based on objective wine properties, that was obtained from actual phisical mesarements.

2 Related Work

In order to move beyond our reliance on human experts in wine market and push forward automation, significant amount of effort has been invested in determining what factors contribute to the subjective quality of wine [1]–[4]. Some work has gone into making objective, quantitative ”electric tongues”, or multisensor potentiometric arrays used to quantify gustatory sensation [5], [6]. Other work has focused more on quantifying certain traits of wine. These are features such as ”savoriness”, or having notes of ”dried fruit” or ”pepper” [7], or perhaps ”balance”, ”flavor intensity” and ”astringency” [8]. All these factors come from objective physical measurements therefore can set the ground as data for automation. Nevertheless, the most difficult, and possibly most fraught goal is to model the overall quality of wine with no additional information other than what can be objectively measured.

In 2009, a group of researchers from Minho in the north of Portugal published a standard dataset in the field of subjective quality prediction and of wine characterization. They collated physical measurements from over 6000 examples of a local, young wine called vinho verde [9], as well as solicited a panel of three experts to judge the quality of these wines on a ten point scale. Their best performing model, an SVM regressor was able to predict quality of red wines within 0.46 points of the judged and within 0.45 points for the white. However, they were only able to achieve this result by manually separating wines by style and training models on each style individually.

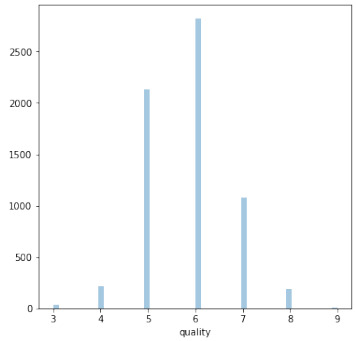

Figure 1: class imbalance

3 Dataset and Features

The dataset considered here includes 6497 examples of vinho verde. It is unevenly split between two styles: 75 % of examples are of white wines and 25% are of reds . Each sample is described by 11 physicochemical features as follows: fixed acidity, volatile acidity, citric acid concentration, residual sugar content, chlorides concentrations, free sulfur dioxide content, total sulfur dioxide content, density, pH, sulphates concentrations, and alcohol content. All 11 features are continuously distributed. Note that each example is also given two labels: its style (red or white), and its subjective rating (integers between 0 and 10). Test set is randomly sampled from dataset in the portion of 20%. The remaining data is randomly split into the training set 80% and validation set 20%. All the samplings are done in a stratified manner. Note that there is strong imbalance in quality classes given in Figure 1 below. Solution to that was binning smaller classes together resulting in 5 classes instead of 7 original.

4 Methods

This chapter describes machine learning models that were used to solve the task of wine quality classification. First basic classification techniques of Logistic Regression and Support Vector Machines (SVM) are discussed. Further LightGBM [10] is considered as an advanced algorithm of gradient boosting in decision trees. Main working principles and highlights of each algorithm are described in subsequent paragraphs.

4.1 Logistic Regression

Logistic Regression is one of the simplest ML models. It is usually chosen as a baseline model to find the metric threshold for other models. In some cases results could be surprisingly high and this model could be preferable because of its simplicity and good performance. The Sklearn implementation of Logistic Regression was used in this paper. We used 'lbfgs' solver and cross-validation technique to avoid overfiting .

4.2 Support Vector Machines

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. It tries to find the hyper-plane in feature space to separate the positive and negative classes. Extensions of SVMs to multi class are also widely used. We are using the one-vs-one scheme.

4.3 LightGBM

Lightgbm is one of the latest tree based gradient boosting models which usually performs significantly better then Random Forests and XGBoost. It is sensitive to hyper-parameter tuning but usually the results are very worth it. This is the main reason that popular algorithm like Random Forests and XGBoost were not tried in this paper. There are many simple methods for hyper-parameter tuning

Table 1: Type Classification

Method	Accuracy
Logistic Regression	0.99

Table 2: Quality Classification

Method	Add. Features	Balanced Acc.	F1
Logistic Regression	no	0.501	0.532
SVM (one-to-one)	no	0.592	0.595
LightGbm	no	0.64	0.649
Logistic Regression	yes	0.464	0.527
SVM (one-to-one)	yes	0.58	0.584
LightGbm	**yes**	**0.663**	**0.668**

like random and greed search but having big parameter space we will use Bayesian Optimization methods for that.

5 Experiments and Results

5.1 Type Classification

This task was rather easy because of good feature separation between 'red' and 'white' classes. Simple logistic regression Gave 99 % accuracy and there was no need to spend more time on this task.

5.2 Quality Classification

Main metrics used for evaluation of discussed models are Balanced Accuracy and weighted F1 score. First note that features were normalized for linear models. Also new features were combined from defaults by creating pairwise multiplications of second order. Last step increased F1 score in some cases significantly.

The best parameters for all models were found using cross-validation on five folds. In case of LightGBM Tree-Structured Parzen Estimators (TPE) algorithm was used as the set of its hyper-parameters is much bigger then the others. Also it is a sample of Bayesian Optimization algorithms which are well known for their superior performance in handling this kind of issues. We used Hyperopt library for that purpose. Accuracy and F1 score could be found in table 2. Per class F1 scores are shown in table 3.

6 Discussion

6.1 Best Model

The model which outperforms all the others is LightGBM. It achieves 0.668 F1 score. Also it's implementation is rather fast compared to other tree algorithms and performs much better then Older

Table 3: Quality F1 scores for each class (top 3 models)

Class	SVM	LightGBM	LightGBM (Add.)	Support
0	0.283	0.383	**0.432**	49
1	0.634	0.693	**0.712**	426
2	0.640	0.671	**0.687**	564
3	0.488	0.543	**0.585**	215
4	0.39	0.384	**0.508**	39

XGBoost And Classical Random Forests. Trainig process is also fast (much faster then SVM) and comparable to logistic regression by sklearn.

7 Conclusion and Future Work

More time could be spent on feature generation and data exploration. One more interesting approach could be using CatBoost Ranker because of ordinal nature of our labels.

References

[1] J. Aleixandre-Tudo, I. Alvarez, M. Garc ıa, V. Lizama, and J. Aleixandre, "Application of multivariate regression methods to predict sensory quality of red wines," Czech Journal of Food Sciences, vol. 33, pp. 217–227, jun 2016.

[2] Y. Gupta, "Selection of important features and predicting wine quality using machine learning techniques," Procedia Computer Science, vol. 125, pp. 305–312, jan 2018.

[3] S. Tempere, S. Per es, A. F. Espinoza, P. Darriet, E. Giraud-H ' eraud, and A. Pons, "Consumer preferences for different red wine styles and repeated exposure effects," Food Quality and Preference, vol. 73, pp. 110–116, apr 2019.

[4] X. Chu, Y. Li, Y. Xie, D. Tian, and W. Mu, "Regional difference analyzing and prediction model building for Chinese wine consumers' sensory preference," British Food Journal, vol. ahead-of-p, oct 2019.

[5] A. Legin, A. Rudnitskaya, L. Lvova, Y. Vlasov, C. Di Natale, and A. D'Amico, "Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception," Analytica Chimica Acta, vol. 484, pp. 33–44, may 2003.

[6] D. Kirsanov, O. Mednova, V. Vietoris, P. A. Kilmartin, and A. Legin,"Towards reliable estimation of an "electronic tongue" predictive ability from PLS regression models in wine analysis," Talanta, vol. 90, pp. 109–116, feb 2012.

[7] J. Niimi, O. Tomic, T. Næs, D. W. Jeffery, S. E. Bastian, and P. K. Boss,"Application of sequential and orthogonalised-partial least squares (SOPLS) regression to predict sensory properties of Cabernet Sauvignon wines from grape chemical composition," Food Chemistry, vol. 256,pp. 195–202, aug 2018.

[8] J. A. Cayuela, B. Puertas, and E. Cantos-Villar, "Assessing wine sensory attributes using Vis/NIR," European Food Research and Technology, vol. 243, pp. 941–953, jun 2017.

[9] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," Decision Support Systems, vol. 47, pp. 547–553, nov 2009.

[10] G. Ke, Q.Meng, T. Finely, T. Wang, W.Chen, W. Ma, Q. Ye, T. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", Advances in Neural Information Processing Systems 30 (NIP 2017)