Predicting Quality Rating of White Wine

Joi Chu-Ketterer and Jon Henin

DSC630

Bellevue University

November 13, 2019

[lchuketterer@my365.bellevue.edu](mailto:lchuketterer@my365.bellevue.edu)

[jhenin@my365.bellevue.edu](mailto:jhenin@my365.bellevue.edu)

Instructor: Becky Deitenbeck

**Predicting Quality Rating of White Wine**

**Abstract**

Wine classification and preference selectivity is difficult to quantify despite the 100-Points wine scoring and other similar systems. To address the issue, this project builds a wine quality prediction algorithm using physicochemical properties of white wine provided by Cortez et al. (2006). Using Cortez’s study as a baseline, Logistic Regression, Random Forest, Support Vector Machine, and Naive Bayes models were created. Random Forest produced similar results to the original model built by Cortex et al. (2006) with an average accuracy of 86.2% and an average precision of 76.3%. Recall statistics were introduced in this study, where Random Forest performed at an average of 53.4% recall. With little significant increase in performance compared to previous models, these models currently show no improvement gains over that of a human expert. However, further modeling and parameter turning could prove useful if paired with ternary classification.

**Introduction**

Wine has been enjoyed by many cultures across the world, and only continues to grow in popularity. An estimation of 246 million hectoliters of wine was consumed worldwide in 2018, with the United States being one of the top twenty countries (Karlsson, 2019). From both a production and consumption standpoint, the quality of the wine is an important factor.

On the production side, wine must meet certain quality standards in order to obtain certifications that are required to sell the wine. The certification requirements are for both a health standpoint, as to assure that the wine is safe to drink, but also from a brand perspective to assure brand standards. These certification processes rely on physicochemical and sensory tests (Ebeler, 1999). While the physicochemical tests are based on characteristics of the wine, the sensory tests are simply taste tests by human experts. Despite experts undergoing extensive training, the subjective scales used for evaluation, such as Wine Spector’s 100-point scale, leave room for inconsistencies. Additionally, critics are unable to review all available wines as the number entering the market continues to increase. According to Wine Spector, it is estimated that critics taste up to 700 wines over just a few days, with over 16,000 wines rated in a year among all critics (Wine Spector). Compared to the 18,5097 uniquely labeled wines in just the US alone this past year, critics are reviewing about 86% of wines available to consumers (Alcohol and Tobacco Tax and Trade Bureau, US).

As wine continues to grow in both popularity and volume, machine learning tools can be used to help determine the quality of wines. This in turn allows wines that may not be tasted by critics to have equal opportunity to thrive in the market. To ensure a good experience for all wine drinkers, it is beneficial to create a platform that standardizes wine ratings.

*Background*

Previous attempts at using a decision support system have been made, yet most have focused on the wine production phase (Ferrer et al., 2008). In most cases, modeling attempts have been made to help classify the type of wine (Vlassides et al., 2001) (Legin et al., 2003) (Moreno et al., 2007) (Yu et al., 2008) rather than the grading of wine. Some of these models were rather successful in categorization, however, due to limited data samples (between 50-170 samples), correlation between physicochemical and sensory panel tests was deemed too difficult a task (Legin et al., 2003). In 2009, Cortez et al., compiled a dataset of 1599 red and 4898 white wine examples and conducted their own predictive study. Using their data and research as a benchmark, this study builds a similar wine grading model using machine learning.

**Methods**

*Data*

The dataset used for this study originated from Cortez et al. (2006) and made public through UCI Machine Learning Repository. The dataset consists of white and red wines; however, only white wine is analyzed within the scope of this project. This results in 4898 observations and eleven physicochemical properties of white wine vinho verde from Portugal between 2004 and 2007. The physicochemical attributes are:

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Predictive variable (based on sensory data):

12 - quality (score between 0 very bad and 10 excellent)

The dataset only includes wines which were going through the certification process. The final wine quality grade was the median value of at least three assessors through blind taste tests.

*Model Evaluation*

In order to maintain the hierarchy of wine quality, binary classification models, similar to Cortez’s study, were created. The quality of wine was grouped into *good* (quality > 6, n = 1054) and *bad* (quality ≤ 6, n = 3816). Binary classification models were chosen for their parallel properties to a wine *passing* or *failing* inspection. Accuracy, precision, and recall statistics were produced to determine the overall performance of each model.

Where *TP* = True positive; *FP* = False positive; *TN* = True negative; *FN* = False negative.

With focus on certification, more importance was put on identifying passing or failing wines. For this reason, Recall statistics were introduced in this study as it was not included in previous ones. Recall is important since it measures how accurately the model predicted the wines that should have passed certification.

*Split Methodology*

Three models with two variations each were created by using two different programing languages, for a total of 12 models. Logistic Regression, Random Forest, and SVM models were created in Python by Henin. Each model was created with and without Principal Component Analysis (PCA) for a total of 6 models within Python. Similarly, Logistic Regression, Random Forest, and Naive Bayes models were created with and without PCA by Chu-Ketterer, for a total of 6 models in R. Principal Component Analysis (PCA) was used to reduce the number of features (n = 4) to determine the benefits, if any. While it might be possible that another value set for PCA might benefit the models, this parameter tuning was not performed. All models were trained and tested using a 10-fold cross-validation method, with 80% of the dataset used for training and the remaining 20% used for testing. The sample split for R and Python were separate from each other. Scoring to determine the best model was done using Accuracy and Recall. Each method explored hyperparameter tuning.

*Python*

The following hyperparameters were chosen for model creation:

Logistic Regression:

· Penalty: *l1* and *l2*

· C: 1, .5, and .1

· Solver: *liblinear*

Random Forest:

· Criterion: *gini* and *entropy*

· Minimum Sample Leaves: 1-10

· Max Depth: 1-10

SVM:

· Kernel: *linear* and *rbf*

· C: 1-10

*R*

The following hyperparameters were chosen for model creation:

Logistic Regression:

· Family: *binomial*

· Maxit: 10, 50, and 100

Random Forest:

· Metric: *accuracy*

· Method: *repeatedcv,* and *cv*

· Number: 2, 10, 20

· Repeats: 3, 10, 15

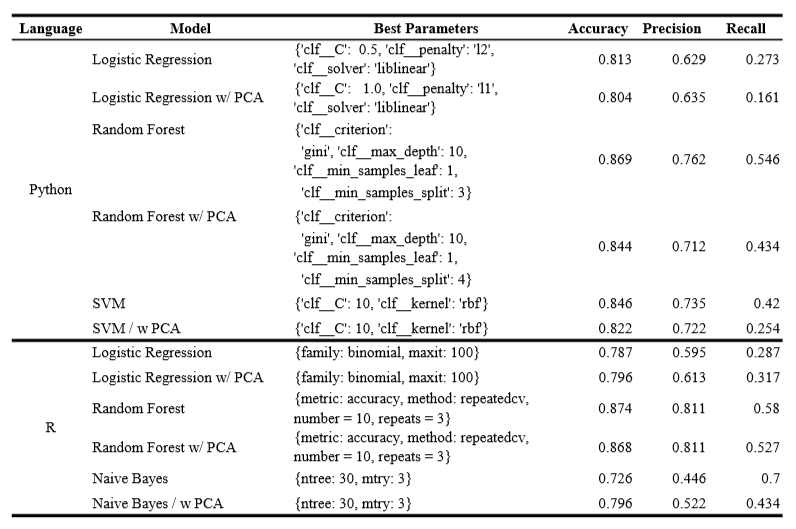
Naive Bayes:

· ntree: *15, 30, 50*

· mtry: 3, 10, 15

**Results**

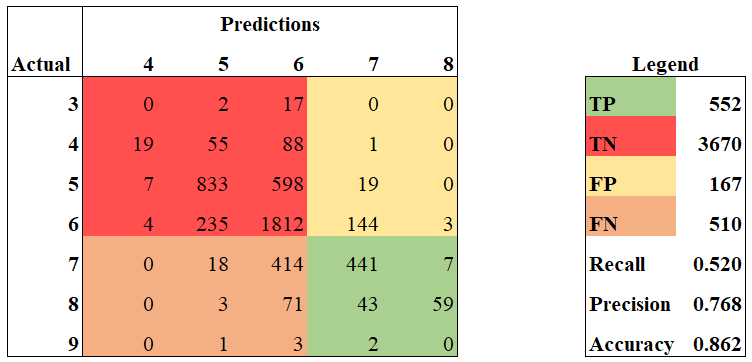
As seen in *Figure 1*, a Random Forest model achieved the highest Accuracy (87.4%) and Precision (81.1%) ratings and performed well in Recall (58.0%) across both Python and R . SVM performed similarly with Accuracy (84.6%) and Precision (73.5%) but saw a drop in Recall (42.0%). Naive Bayes had the highest Recall (70.0%), did well in Accuracy (72.6%) but suffered in Precision (44.6%). Logistic Regression performed poorly compared to all other models. Principal Component Analysis was not found to improve any model.

**

***Figure 1.*** *Model Results of the best Hyperparameters*

**Conclusion**

Precision, Recall, and Accuracy are the three metrics used to analyze the performance of the models. In addition, a performance comparison was made to Cortez’s study. However, since Cortez conducted a regression analysis, predicting values along the 0-10 score range, and this study conducted a binary classification analysis, Cortez’s predictions results were converted into the binary classifications of this study; *good* and *bad* wines.



***Figure 2.*** *Original Study Regression predictions color coded to our binary classification.*

Binary classification was chosen for this study to best mimic the binary certification process of *passing* or *failing*. It is made aware that the chosen classification boundary of 6 on the quality score scale is arbitrary. While this criterion might hold for certain wines, it might not hold true for wines held to a higher standard.

The other important thing to note is that the original study reported precision calculations as model accuracy. Models were achieving a precision of 64.6% for white wines when set to a tolerance of 0.5 (±1 score values were considered correct) and precision around 90% when set to a tolerance of 1 (±2 score values were considered correct). When converting Cortez’s regression predictions to this study’s binary system, Cortez’s predicted outcomes had a Recall value of 51.9% and a Precision value of 76.7%. Comparatively the best performing model in this study, a Random Forest model, scored a Recall of 58.0% and Precision 81.1%, making this new model slightly more accurate at predicting good wines than the original study.

Recalling that the Random Forest model had a 58.0% recall and the number of *good* wines in the dataset was 1054, this equates to about 443 wines. Considering each wine sample could represent an entire production, rejection of a wine could be very costly for a winery. When taking into account Precision at 81.1%, this allows *bad* wines to get classified as *good* wines 1 in 5 times. This could potentially have implications such as damage to the brand and/or health risks.

Given that the original study’s results showed higher accuracy and precision toward the extremes, a ternary classification may perform better than the binary classification used in this study. A ternary classification could help both really good wines and really bad wines get identified with higher confidence while giving more latitude to middle quality wines. The higher confidence levels would negate the need for human review for excellent and poor wines. This in turn would reduce the number of wines human critics would need to review and hopefully allow all wines on the market to receive a rating, rather than just 86%. Given this proposed function, the models created in this study would strictly be a decision support tool rather than a replacement of human experts.

In future analysis, another aspect to research would be how well the models from this study would perform outside the specific region of the wine dataset. The dataset used is comprised of samples from only white wine vinho verde from the Minho region of Portugal, collected between 2004 and 2007. Considering that grapes and types of wines can be geographically specific, it is most likely that the physicochemical properties of the wine are also geographically specific. If a certification process is done locally, then it makes sense to develop a model based on geographical boundaries. However, if the certification process is more globally managed, then geographical boundaries might not be as important, and a robust dataset featuring many different wines would be needed to develop the necessary model.

**Acknowledgements**

Henin and Chu-Ketterer would like to thank their peers and professors in DSC630. This work was made possible with the resources provided by Bellevue University.

**References**

A. Asuncion and D. Newman. UCI Machine Learning Repository, University of California, Irvine, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 2007

A. Legin, A. Rudnitskaya, L. Luvova, Y. Vlasov, C. Natale, and A. D’Amico. (2003). Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception. Analytica Chimica Acta, 484(1):33–34, 2003.

À. Nebot, F. Mugica, A. Escobet. (2015, July). Modeling Wine Preferences from Physicochemical Properties using Fuzzy Techniques. In *SIMULTECH* (pp. 501-507).

R. Arther. (2019, April 15). Global wine production reaches record level. Retrieved from<https://www.beveragedaily.com/Article/2019/04/15/Global-wine-production-reaches-record-level>

D. David. (2013, Jun 22) Wine-tasting: it's junk science. Retrieved from <https://www.theguardian.com/lifeandstyle/2013/jun/23/wine-tasting-junk-science-analysis>

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.D. Smith and R. Margolskee (2006). Making sense of taste. *Scientific American*, Special issue, 16(3):84–92

H. Yu, H. Lin, H. Xu, Y. Ying, B. Li, and X. Pan. (2008). Prediction of Enological Parameters and Discrimination of Rice Wine Age Using Least-Squares Support Vector Machines and Near Infrared Spectroscopy. Agricultural and Food Chemistry, 56(2):307–313

I. Moreno, D. Gonz´alez-Weller, V. Gutierrez, M. Marino, A. Came´an, a. Gonz´alez, and A. Hardisson. (2007) Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks. Talanta, 72(1):263–268.

J. Ferrer, A. MacCawley, S. Maturana, S. Toloza, and J. Vera. (2008). An optimization approach for scheduling wines grape harvest operations. International Journal of Production Economics, 112(2):985–999.

K., Per. (2019, Apr 14) World wine production reaches record level in 2018, consumption is stable. Retrieved from<https://www.bkwine.com/features/more/world-wine-production-reaches-record-level-2018-consumption-stable/>

R. Susan. (2016). Millennials Drink More Wine Than Any Generation*.* Retrieved from <https://www.forbes.com/sites/thecut/2016/03/08/millennials-drink-more-wine-than-any-generation/#4c3c261175ea>

S. Ebeler. (1999). *Flavor Chemistry - Thirty Years of Progress, chapter Linking flavour chemistry to sensory analysis of wine*, pages 409–422. Kluwer Academic Publishers

S. Vlassides, J. Ferrier, and D. Block. (2001). Using Historical Data for Bioprocess Optimization: Modeling Wine Characteristics Using Artificial Neural Networks and Archived Process Information. Biotechnology and Bioengineering, 73(1).

United States of America. Alcohol and Tobacco Tax and Trade Bureau. *COLA Registry.* TTB F 5100.31: Application For and Certification/ Exemption of Label/Bottle Approval.

Wine Institute (2019, Jun 24). Wine Consumption in the U.S. Retrieved from<https://www.wineinstitute.org/resources/statistics/article86>

Wine Spectator (n.d.). About Our Tastings. Retrieved from<https://www.winespectator.com/articles/about-our-tastings>