ASSESSING WINE QUALITY BASED ON PHYSICOCHEMICAL PROPERTIES

Ian Konigsberg, Maximilian de Ledebur, Elizabeth Lentine

ORIE 4741: LEARNING WITH BIG MESSY DATA
CORNELL UNIVERSITY, FALL 2019
PROFESSOR MADELEINE UDELL

December 11, 2019

CONTENTS

I. Introduction

Winemaking with the Vitis Vinifera grape is a practice that has deep cultural and economic roots reaching back to the inception of civilization. In its current state, the industry relies on the expert opinion of oenologists, or sommeliers, to decide which wines stand out as superior. Since this is inherently a subjective evaluation, our team set out to predict the quality of wine in order to see if machine learning techniques could help the wine industry develop and market only the finest wines. This is a preliminary study used to determine the feasibility of such a task and to decide whether extensive data collection is worthwhile. Our models make use of a publicly available dataset on the physicochemical properties of the Portugese wines named Vinho Verde (”veeng-yo vaird”), which translates to ”green wine”. This unique name is not a reference to the color of the wine or grapes, but rather the fact that the wines are released a few short months (3-6) after harvest. Located in the Northwestern area of Portugal, the Vinho Verde region is one of the oldest and largest winemaking regions in Portugal. Vinho Verde wines are globally recognized for their many distinct qualities, including: a refreshing, fruity flavor, a great pairing with salad and seafood, and a reasonable price. Most Vinho Verde varieties are white wines, but some vintners also produce red and pink varieties [5]. Due to this disparity, the bulk of the data used in our study came from white wines. We hope that the optimistic results of this study will encourage vintners around the world to invest in further data collection of physicochemical properties in order to employ similar techniques to verify the quality of their wines.

II. Problem Description

This study will make best use of the previously mentioned Vinho Verde dataset to predict the qualities of each distinct wine. This dataset includes the wine type (red or white), several physicochemical properties, and the quality of the wine, as assessed by oenologists. We will give a detailed description of the dataset in the proceeding section. The massive global wine industry impacts a lot of people, from grape farmers to retailers and consumers. The value of the global industry is estimated to be well over $300 million [2]. Winemakers produced 293 million hectolitres in 2018 alone [3]; to put this in perspective, this is roughly 12,000 Olympic-sized swimming pools of wine!

Although many established vintners have ample funds to spend on research and development, there are many individuals and small producers producing beverages for hobby, personal consumption, or as a small business. These important players in the winemaking community may not be able to afford these expenditures and rely on limited resources, such as family and friends, to taste and rate their wines. A successful machine learning project would be able to greatly improve accessibility of development tools for small producers and reduce costs for large firms. If we can show that simple algorithmic tools can consistently and confidently assess wine quality, the total investment from winemakers would only depend on the data collection of their wines physicochemical properties. In fact, we will later show that only a small number of variables can predict wine quality with high accuracy. If it were to turn out that such a tool can be created to assess any variety of wine, this information would have the potential to greatly disrupt an industry that is currently dominated by select human experts.

While our team is interested in properly predicting the quality of each wine on an ordinal scale from 1 to 10, we are focused on identifying low and high quality wines with nearly perfect precision. Following similar analyses, winemakers can swiftly identify their best wines and techniques without needing to spend money and effort on sampling opinions. In order to capture a general idea of the quality of a wine, we have grouped the dataset into bad, okay, and good. As a result, we understand our project as primarily a classification task and also a regression problem on the full quality scale of 1 to 10.

III. Description of Data Used

The Vinho Verde dataset includes 6,497 wines, which are either red or white. Because of inherent differences in red and white wines (which we will soon explore), our team decided to first perform analyses on the red and white wines separately and later on the entire dataset. The featurespace

of the dataset includes the wine color and 11 additional physicochemical properties of the wine: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol.

The output variable that we wish to predict is the quality of wine on a scale of 1-10. The wine quality in the dataset was determined by the median value of three oenologists' ratings on a scale of 1-10. Figure 1 shows the distribution of wine qualities for both reds and whites.

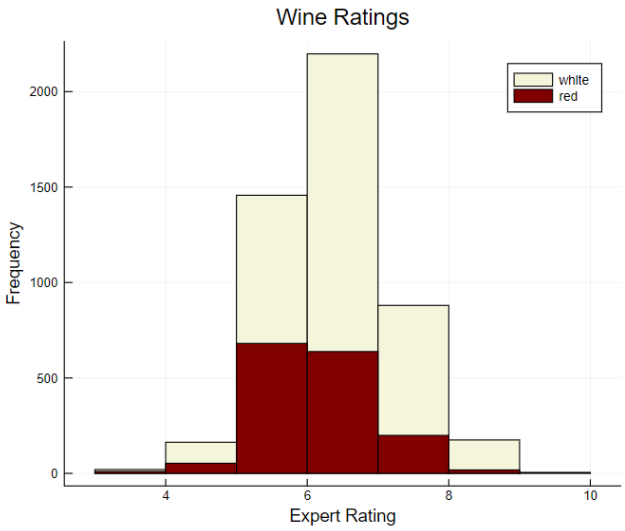

Fig. 1. Distribution of Wine Ratings by Color

As previously mentioned, our teams primary focus is to identify bad and good wines with very high precision. We have decided to categorize the wines as such: good is a score of 7 or higher, bad is a score of 4 or lower and okay is a score of 5 or 6. Using these designations, our dataset has a total of 246 bad wines, 4974 okay wines and 1277 good wines. Figure 2 shows the distribution of wines in each category.

The dataset we chose was attractive for our project because it was collected by the University of Minho, a prestigious school in the Vinho Verde region, and the Viticulture Commission of the Vinho Verde Region (CVRVV). As a result, our team is confident that we can trust the accuracy of the figures within the dataset. In addition, we were happy to find that the dataset included very few missing values; there were only 34 incomplete rows, meaning only 0.52% of our rows had any

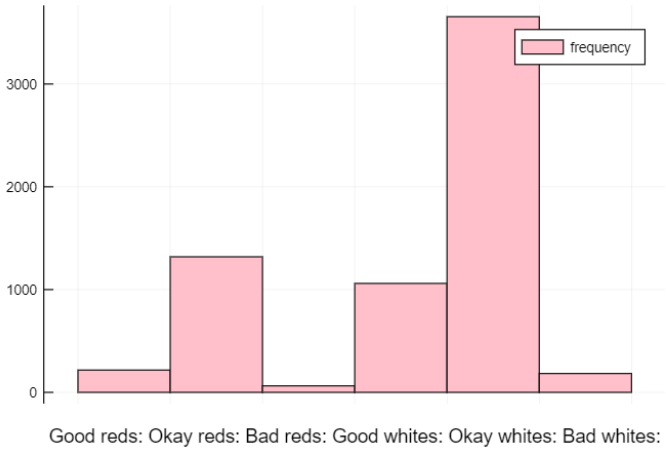

Fig. 2. Distribution of Wine Classes by Color

missing entries. We decided to drop these incomplete rows after imputing the missing values and observing no difference in model performance.

IV. Exploratory Data Analysis

In order to demonstrate that the physicochemical properties of a wine impact its quality and justify building a model based on these predictors, we visualize some examples of the trends displayed when viewing measurable properties of wine against eventual quality assessments. In particular, we noticed that while some variables seemingly have little to no effect on the quality, other properties such as volatile acidity, citric acid, alcohol and sulphates have a clear impact on quality. Figure 3 contains boxplots of these properties against wine quality and reveals clear trends. One example lies in the relationship between volatile acidity and quality of red wines; well-rated varieties were much more likely to show low levels of volatile acidity.

To elaborate on our reasoning for bisecting our data by color, let us examine the box plots for citric acid vs. quality of both red and white wines. We can see from Figure 3 that as the quality of a red wine increases, so does its average citric acid content. However, for white wines, there is not nearly as much of a difference in citric acid content across different qualities and there are significantly more outliers.

Accordingly, surface-level knowledge of oenology implies that red and white wines taste, smell, appear, and are assessed quite differently because

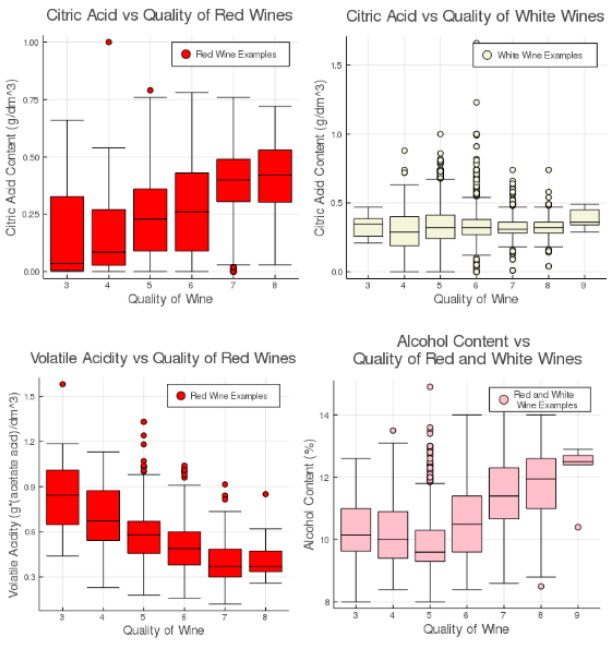

Fig. 3. Boxplots

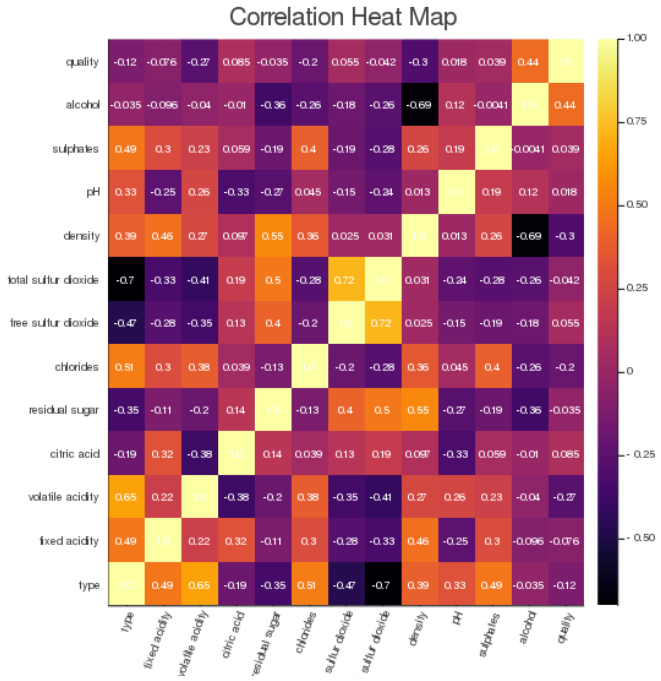

Fig. 4. Correlations of Variables

of the fundamental differences in the techniques used in making them. Namely, red wines are fermented with the seeds and skins still on the grapes. It is not hard to see that such a vast difference in fermentation technique could lead to distinctive differences in the chemistry of the end results.

The correlation heat map in Figure 4 demonstrates the range of correlation between the properties in our data. This is useful to us for a few reasons. Noticing strongly correlated properties could allow us to drop variables whose effects on our prediction are already being represented strongly enough by those properties with which they are highly correlated. The purpose for dropping a variable is two-fold; it will reduce the complexity of every model and therefore the risk of overfitting, a phenomenon which we will elaborate on in our description of the models we used. In addition, reducing complexity is beneficial to the end user of the model as it reduces the amount of work and costs associated with data collection.

Let us examine the strong negative correlation between density and alcohol content visible in the heat map. Alcohol is not as dense as other chemicals found in wine, so we confirm with our intuition that more alcohol in a blend would yield less dense wine. This leads us to consider if either alcohol or density is a variable worth dropping.

V. Regression Modeling Approaches

A. *Least-Squares Regression*

We began with simplistic models to understand the importance of each feature and set a baseline for us to move forward. Our very first attempt was a least-squares linear regression model. The regression was fit to predict the quality of wine between 1 and 10 using all features in the dataset. This gave us a good idea of a rough lower bound for the accuracy scores we should be looking for. As a least-squares model can be thought of as a line of best fit, this regression sought to minimize the squared differences between the predicted and actual wine qualities. Given that our dataset includes 12 features, it would be impossible to visualize this line in the proper dimensions. As seen in Figure 5, the linear model does a poor job, as a successful implementation would show

plot points along the diagonal on the plot. Notice that the predicted quality scores are continuous whereas the actual scores are only discrete, integer values.

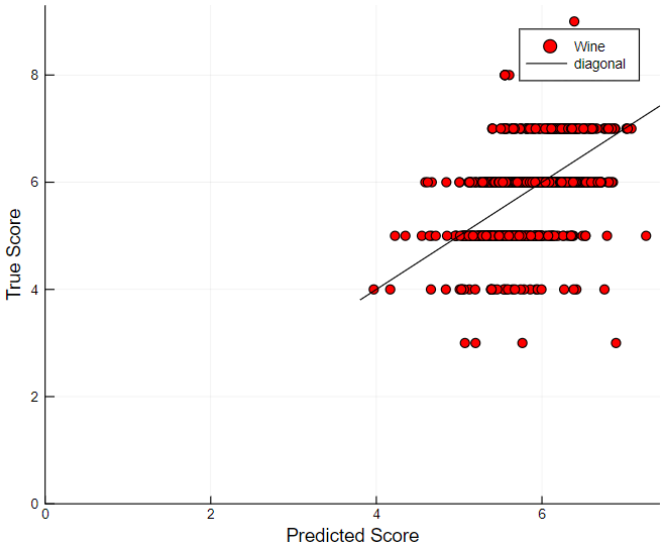

Fig. 5. Least-Squares Regression

Our results for mean squared error and mean absolute error were as follows:

White mean squared error, test: 0.5820
Red mean squared error, test: 0.3940
White mean squared error, train: 0.5666
Red mean squared error, train: 0.4223
White mean absolute error, test: 0.7629
Red mean absolute error, test: 0.6277
White mean absolute error, train: 0.7527
Red mean absolute error, train: 0.6498

B. *Quantile Regression*

Now that we had a baseline as described, we began a quantile regression to locate the most important explanatory variables. More specifically, we were most interested in which explanatory variables would greatly affect the quality of the higher quality wines, in alignment with the overall goal of our study. We found this model to be an appropriate choice since it decides a line of best fit at a specified quantile.

Aside: Standardizing Data *Before exploring the results and interpretation of our quantile regression, we must address our decision to standardize our data. Standardizing our data puts all of our explanatory variables on the same scale. We quickly noticed that our particular dataset exhibited large differences in magnitude of each explanatory variable due to variation in units. For example, total sulfur dioxide is typically a number greater than 100 whereas chlorides is typically a number less than 0.1. Standardizing the data significantly increases interpretability for each model because it standardizes the weights. We will decide which features are most important in determining the quality by looking at the weights with the highest absolute value.*

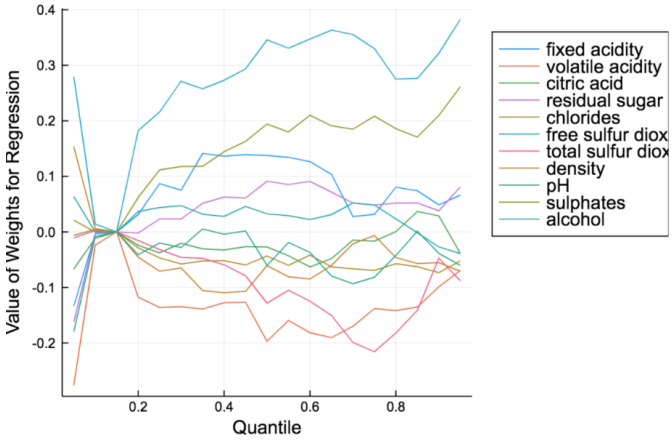

Fig. 6. Quantile Regression Plot

The results of the quantile regression model are displayed in Figure 6, with the regression quantile on the x-axis and the value of the weights on the y-axis. These weights can be thought of as the slope of our line of best fit in each dimension; therefore, a large positive weight for a variable implies that this variable has a large positive impact on the quality of wine. We pay special attention to the right side of the graph because the highest quality wines are in the upper quantiles. The graph tells us that, of all the explanatory variables, alcohol content and sulphates have the greatest impact on wine quality for those higher quality wines. In fact, alcohol content seems to be the strongest explanatory variable at all quantiles. This news is supportive of our findings in the exploratory dataset analysis portion of this paper (recall the previous box plot showing the positive correlation between alcohol content and wine quality). In the following models, we will use the information

provided by this plot to make informed decisions regarding which variables to use.

C. Further Feature Reduction

We performed one final regression to support the claims made by the quantile regression plot. This regression is known as Lasso Regression and is commonly used to produce sparse results in the weights vector. Employing lasso on our data added a regularizer that promoted sparsity by penalizing weight vectors with entries of large magnitude. Since we wanted to confirm the results of our prior feature selection, we altered the tuning parameter of the lasso regression in order to force the weight vector to contain only three non-zero entries. We can then interpret this 3-sparse weight vector result to contain nonzero entries corresponding to the explanatory variables that the lasso method found to be the most significant predictors of quality; results are summarized in Figure 7.

Explanatory Variable	Value of Weight
Volatile Acidity	-0.17979
Sulphates	0.00364
Alcohol	0.27392
All other variables	0

Fig. 7. Lasso Regression Weights

This was a reassuring result; the consensus across our methods is that volatile acidity, sulphates and alcohol are the three strongest predictors of wine quality. Notice that alcohol and sulphates have (on average) the largest weights across all quantiles while volatile acidity has (on average) the most negative weight across all quantiles. As we move forward we pay special attention to these three variables.

VI. Classification Modeling Approaches

As we tried different models on our data, we chose to identify two specific performance metrics for proper comparison across techniques. Ultimately, we decided on precision and recall, with a slight emphasis on precision. For the purpose of our report, let us define these metrics as follows:

Precision: For the number of times the model guessed class i, what percentage were actually class i?

Recall: For the number of class i points in the dataset, what percentage did the model predict correctly?

Maximizing our recall score ensures that a good wine, when run through our model, will be assessed as good. Maximizing our precision score ensures that any wine our model classifies as good actually has a true rating of at least 7. We feel that in practice, the most important use for our model will be to determine what wines are of objectively good quality for their variety and use this to decide whether a wine is released and perhaps how it is priced. It would be highly undesirable to overestimate the quality of a wine and tarnish the reputation of a vintners ability to produce high-end products; thus, we are particularly motivated by our model's precision. Alternatively, we do not want to mislead a vintner by mistaking a good wine for a not good one. However, we feel that even if a formula is predicted to be of lower quality and is released in smaller quantities or at a lower price, the observed popularity among consumers could allow recovery from such a misclassification. We used this logic to justify emphasizing precision over recall.

This point in our modeling marks a large transition in the project. We have gained many insights about our explanatory variables and now move on to our primary goal, the classification problem. In the subsequent models, we have decided to focus on learning the good wines. The same logic stated earlier applies: a wine of quality of 7 or higher is designated as good, otherwise the wine is designated as not good.

A. Balancing the Data

We have an unbalanced dataset with many more not good wines (0) than good wines (1). To prevent overfitting to the 0 class, we decided to test how including class weights would affect the overall performance of our model. In doing this, we hope that the larger penalty for misclassifying "good" wines will reduce our model's bias towards the majority class ("not good" wines) and improve its ability to correctly classify "good" wines. We

chose logistic regression as a method for doing this as it is a simple classification model that can demonstrate how these weights would significantly impact the outcome of our testing scores.

Precision for both models were fairly similar, as seen in Figure 8, but when class weights are included, the percentage of good wines identified doubled in both sets. Since we are most interested in correctly classifying "good" wines, we chose to incorporate class weights in all future models to combat the unbalanced nature of our dataset.

B. Simple Decision Tree

An easy-to-understand tool we introduce to explore our classification problem is the decision tree. A decision tree works by maximizing the uniformity of the final groupings displayed on the right side of the graph in Figure 9. The tree determines and subsequently asks the most important questions. Interestingly, the majority of questions in our decision tree were regarding sulphates and alcohol, cementing the idea that these are the strongest explanatory variables. Each wine example will move through the decision tree answering each question until it is eventually placed into a group with wine examples that have answered the same series of questions in the exact same manner. The hope is that at the end of the decision tree each grouping of wine examples will be comprised of either mostly good wines or mostly not good wines. We visualize our decision tree in Figure 9.

We established this notion of a decision tree as a stepping stone towards our next model, a random forest. As the name suggests, this is a collection of decision trees and provides an aggregate result across them. We do not linger on the single tree, as the results are generally poor. While some final groupings of wines are all of the same quality class, many final groupings are mixed and therefore this single decision tree is not a useful model. Instead, we will make use of many iterations of such decision trees in a random forest model to encounter some of the best results we have seen yet.

C. Random Forest

As promised, we extended the idea of the decision tree to a random forest model and used 10-fold validation to watch the accuracy scores

Without Class Weights	**Precision**	**Recall**
Training Data:		
Not Good Wines (0)	0.84	0.95
Good Wines (1)	0.58	0.26
Accuracy Macro Avg	0.71	0.61
Accuracy Weighted Avg	0.79	0.82
Testing Data:		
Not Good Wines (0)	0.84	0.96
Good Wines (1)	0.59	0.26
Accuracy Macro Avg	0.72	0.61
Accuracy Weighted Avg	0.79	0.82
Class Weights 0 = 1, 1 = 2	**Precision**	**Recall**
Training Data:		
Not Good Wines (0)	0.88	0.86
Good Wines (1)	0.49	0.53
Accuracy Macro Avg	0.68	0.70
Accuracy Weighted Avg	0.80	0.80
Testing Data:		
Not Good Wines (0)	0.88	0.87
Good Wines (1)	0.49	0.52
Accuracy Macro Avg	0.69	0.70
Accuracy Weighted Avg	0.81	0.80

Fig. 8. Performance of Logistic Regression, Weighted vs. Unweighted

Fig. 9. Decision Tree

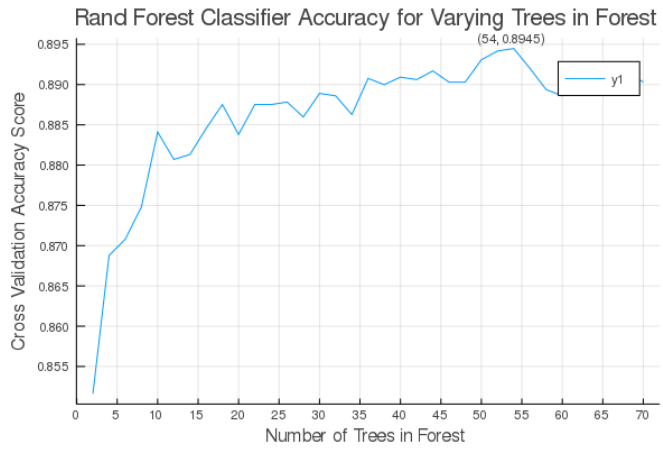

Fig. 10. Random Forest Accuracy by Number of Trees

evolve with the number of trees in the forest. As previously seen in the logistic regression model, adding class weights to the random forest model increased our recall score. To obtain the same benefit, we included them here as well. Figure 10 demonstrates a sharp increase in accuracy past the first few decision trees and somewhat of a convergence around 89% accuracy for many trees.

Testing

	precision	recall	f1-score	support
0	0.91	0.97	0.94	1043
1	0.83	0.58	0.68	250
accuracy			0.90	1293
macro avg	0.87	0.78	0.81	1293
weighted avg	0.89	0.90	0.89	1293

Training

	precision	recall	f1-score	support
0	1.00	1.00	1.00	4149
1	1.00	1.00	1.00	1021
accuracy			1.00	5170
macro avg	1.00	1.00	1.00	5170
weighted avg	1.00	1.00	1.00	5170

Fig. 11. Random Forest Performance

The random forest appears to overfit the training set slightly as seen in Figure 11, but it gives the best precision score for good quality wines out of any of our attempts up to this point. This means that this model is the most reliable when it predicts a wine to be good.

VII. Results and Implications

A. *Assessment of Results*

We validated over several different techniques that sulphates, alcohol content, and volatile acidity are the most relevant predictors of a good Vinho Verde wine of any color. Our results suggest that since quality can be estimated with high accuracy in this manner, good Vinho Verde wines should exhibit these same patterns in their physicochemical properties. Since we only used one regional variety and less than 7,000 wines, we cannot confidently say that our analysis extends across the entirety of oenology. Optimistically, we believe that with similar procedures, one would be able to learn the chemical markers of a quality wine in other varieties as well, herein lying what we deem to be an important proof of concept about winemaking.

To reiterate our findings for the Vinho Verde, both classification and regression showed:

- A strong positive relationship between alcohol content and wine quality
- A negative relationship between volatile acidity and wine quality
- A small positive relationship between sulphates and wine quality

B. Fairness and Destructive Potential

Our model does not involve any protected attribute that is regulated by federal law to enforce fairness. Since these features are chemical measurements, they would not be tied to a protected attribute or group either. The outcomes of our model could affect business decisions, but our group does not believe at this time that scoring wines would adversely affect any group of individuals.

However, we would like to acknowledge that any machine learning performed on the existing data for wine scores is subject to the inherent biases of the sommeliers whose scores create the aggregate rating represented as the "true" rating of a wine. Sommeliers undergo extensive training so they can be as objective as possible, but each score is subject to fluctuation in preference and human sensory ability. In this way, training an algorithm on these subjective outputs has the potential to continually reinforce biases present in the original data, failing to represent a true rating. Although we do not think this constitutes labeling our findings as a Weapon of Math Destruction [4], we wish to point out the similar effect of human bias on the results of our model.

C. Conclusion

We were very happy to find that physicochemical properties of wine give meaningful insights into the quality of wine; thus, vintners should consider fine-tuning their techniques with the assistance of machine learning. We strongly encourage vintners from around the globe to invest in the collection of the physicochemical properties of their wines in order to perform similar analyses.

References

[1] 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.

[2] Analyzing the Global Wine Industry, 2019 to 2023 - Expected to Cross $420 Billion by the End of 2023 - ResearchAndMarkets.com.. Business Wire, 19 Mar. 2019

[3] Karlsson, Britt. Record Global Wine Harvest In 2018, Stable Consumption. Forbes, 14 Apr. 2019.

[4] O'Neil, Cathy. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Penguin Books, 2017.

[5] Signer, Rachel. "7 Things You Need To Know About Vinho Verde — VinePair." Drinking Is Culture - Learn About Wine, Beer Spirits — VinePair. VinePair, 27 May 2016. Web.