Wine Quality Prediction

Kai Yu

Weiye Yao

Supeng Huang

Introduction

● Background

The wine industry, a field rooted in tradition, has often been challenged by the subjectivity of consumer preferences and the ambiguity of factors contributing to a high-quality wine. However, in the rapidly evolving digital age, predictive analytics has become a more powerful tool. This report presents the findings of our empirical study that aimed to use the power of predictive analytics to understand the complex chemistry of wine quality.

In an increasingly competitive global market, the quest for quality has become a key differentiator for wine producers. As the industry's revenues soared to US$309 billion in 2022 and projected to grow at a rate of 5.52% annually over the next five years, there is a clear incentive for producers to leverage every tool at their disposal to increase their market share [1]. Yet, despite the potential, only a handful of wine varieties have managed to do so.

Our motivation was to address this gap, using predictive analytics to link the complex chemical attributes of wine to the elusive notion of quality, as perceived by consumers. We sought to illuminate the key factors that drive consumer opinions, providing a roadmap for producers to improve their product quality and help them capture a larger share of this lucrative market.

The study was designed to address three key research questions: how to predict the subjective quality score of wine using quantitative chemical attributes, which specific chemical attributes strongly correlate with quality scores, and how we can leverage this predictive model to enhance the subjective quality of wine. The ultimate objective was to equip wine producers with insights that could enable them to refine their production processes, creating superior wines that resonate with consumers.

● Data description

The dataset utilized in our study comes from the University of California, Irvine machine learning repository, collected over a span of three years (2004-2007 [2]. It specifically focuses on the 'Vinho Verde' red wine variant from Portugal. Each wine was evaluated in a laboratory and a quality score was determined based on the median of at least three blind sensory tests conducted by wine experts [2]. Our task is essentially using the wine's physicochemical attributes quantitatively to explain human sensory analysis.

The dataset includes a detailed examination of 11 chemical characteristics of each wine, along with a quality score that ranges from 1 (very bad) to 10 (excellent). It encompasses 1,200 different wines with no missing data, offering a robust foundation for our analysis. The wine quality score is used as the dependent variable in this study, while the 11 physicochemical attributes are the independent variables. These attributes span various measurements,

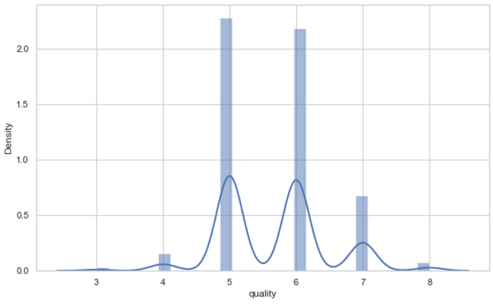
including alcohol percentage, volatile and fixed acidity, residual sugar, density, and quantities of sulfur dioxide, chloride, salt, and sulfates. The detailed explanation of its chemical properties is in the appendix.

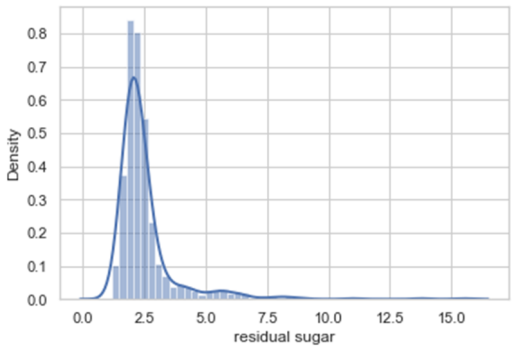
This report is structured into four comprehensive sections. Section 1 delved into an exploratory data analysis, facilitating a deeper understanding of the wine data's distribution and identifying appropriate modeling frameworks. In Section 2, we explained the model employed in our study, and measured the results against potential alternatives, thereby providing insights into potential ways for model enhancement. Section 3 conducted an in-depth examination of the results, utilizing metrics to deliver a robust analysis. Finally, in Section 4, we draw upon our findings to propose model real-world suggestions that could be instrumental in future wine production.

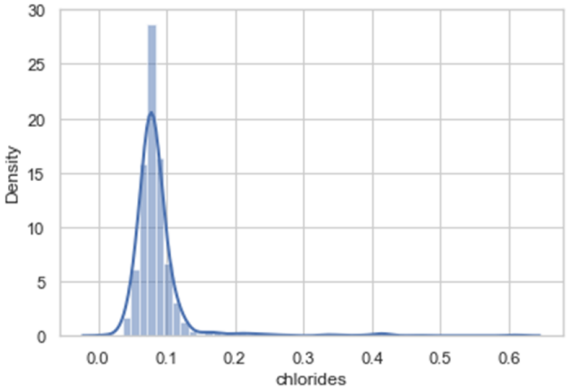
Exploratory data analysis

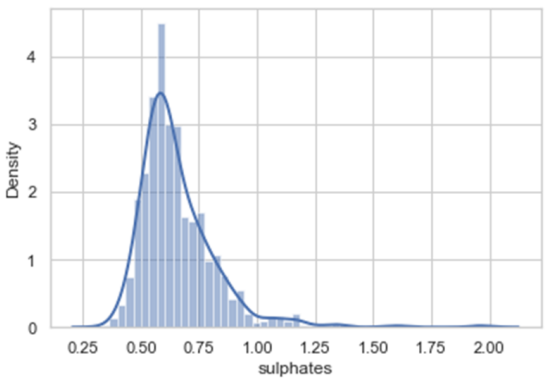
- ## Distribution statistics

The dataset in our project are all continuous variables, so we want to use various data visualization techniques to explore the distribution of them by proper density plots. In our first step of exploratory data analysis, we identified the distribution of quality scores. A histogram of the quality scores, shown in Figure 1, shows that most of the quality scores fall within 5 and 7, indicating the general average quality of produced wines. The lack of uniform distribution in the values might skew the outcomes, as no grouping of features can yield either ideal or inferior wines. To counteract this issue, we suggest implementing the Synthetic Minority Over-sampling Technique in later modeling. This method partitions the sample datasets into several training sets, thus improving the integrity of the analysis. For other features distribution, we highlight histograms of a few typical variables: residual sugar, chlorides, and sulfates. The graphics reveal a skewed distribution in need of feature engineering, evident for log transformation in later standardization for better regression results.



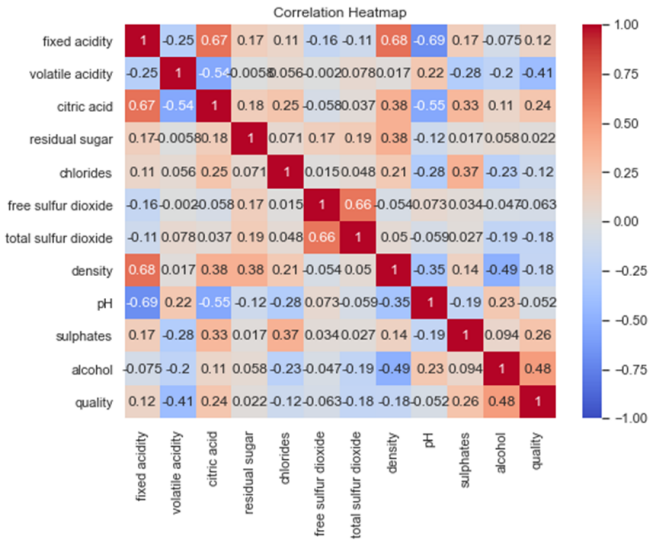




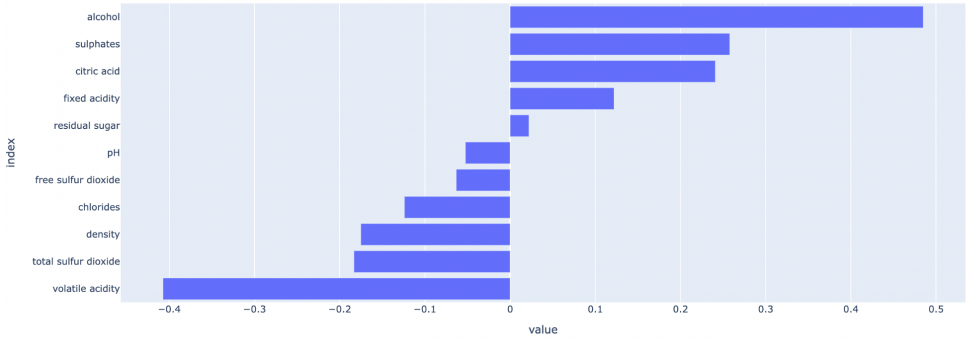


- Correlation analysis

Following this, we plot a correlation matrix of different features, with a heatmap indicating the strength of correlation. Our analysis reveals that certain attributes, specifically chlorides and sulfates, fixed acidity and pH, fixed density and density, and acidity-related characteristics (including fixed, volatile, and citric acid), exhibit significant correlation.



In a more visualized way, we plotted the relationship between quality score between features values. The alcohol and the volatile acidity are two of the most important factors influencing wine's quality according to our data. This provides insights about which specific chemical attributes strongly correlate with quality score and can be further validated by using a predicted model with critical variables intensified.



To dig into the correlation between acidity and quality score, for example, we here depicted a scatter plot chart with superior quality heighted. We find that the volatile acidity and the citric acid have a negative correlation as shown in the heatmap and the wine with a high quality rate tends to have a higher citric acid level and a lower volatile acidity level. Most of the higher-quality wines are grouped together on the right-bottom side of the plot. From this perspective, we are more aware of the combined effects of two independent variables that give rise to higher quality wine.

Model construction

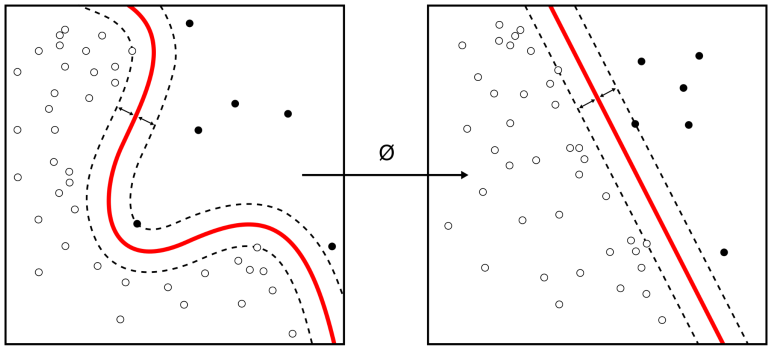
- ## Model description

To establish the predictive model, we observed that the predicted variables are of ordinal values and assuming the order of wine quality is not linearly elevated, for this reason we want to see the different performance of classification models. In our selected models, they include random forest, decision tree, k-nearest neighbors, logistic regression and support vector machines.

In random forest and decision tree, they are of similar type in classification task. A decision tree is a supervised machine learning algorithm mostly used for classification by partitioning the data space into distinct non-overlapping regions. The decision tree constructs these regions by splitting the features of the data under different thresholds and the process continues recursively. The predicted result is then determined by the mode of the training observations of the regions they belong to. This method is easy to interpret based on the trees plot. Random Forest is an ensemble learning method, building upon the decision tree concept. It creates a multitude of decision trees and is grown on a bootstrapped sample of the feature space during the training process. In this case, the prediction output is the class that represents the mode of the classes of the individual trees. This method is less interpretable but normally with better prediction results because the two sources of randomness help to decrease the variance of the model without increasing the bias.

KNN is also a supervised learning algorithm that can easily handle multi-class classification. It performs the prediction by measuring the distance of new observation against all identified instances from the training dataset and selecting the nearest class as its prediction result. The disadvantage is that it can be sensitive to irrelevant features and the scale of the data.

Logistic regression is based on traditional linear classification methods and adds a logistic function to predict the outcome in terms of the likelihood of belonging to a particular class. It provides a probabilistic view of class membership and easily extends to multi-classification problems. While for SVC, it is a more optimized version of basic linear classification, finding a hyperplane that best divides a dataset into two, the one that represents the largest separation between two classes. It is more powerful in our nonlinear prediction case as it can implement kernel function on features to conduct linear classification but it is hard to find a suitable one.



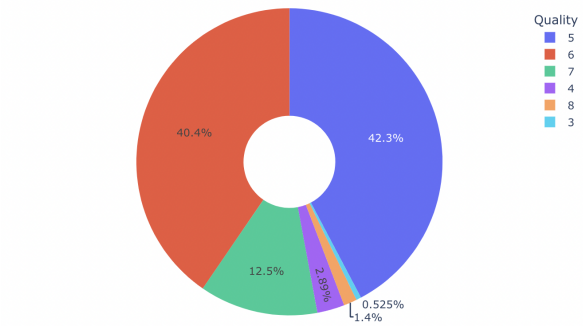
- Baseline performance

We performed the above classification models on our raw wine dataset. The baseline result comes to show that the random forest model with max depth 15 had the best performance on our original data and the KNN and the SVC model performances are not satisfactory, whose cross-validation accuracies were below 50% as these two models are not scaled well due to data variation.

- Approaches of Improvement
    - Standardization

The SVM and the KNN model performance improved significantly, with the accuracy of the SVC model performance improving around 10% after normalizing the data. However, overall, the performance of the current models are not achieving 70%. For this reason, we chose to employ an oversampling strategy to deal with the problem of data imbalance.

    - Oversampling



We find that more than 90% of data are in group 5,6,7 as observed in exploratory data analysis. We try to train similar amounts of data in each group by SMOTE(Synthetic Minority Over-sampling Technique), which creates synthetic (not duplicate) samples that are close in the feature space of the minority class in order to achieve a balance between the minority and majority classes. During the resampling, we make each group contain 160 data points. After oversampling, the performance of all models except for logistic regression improved dramatically. The reason might be that the data is not linear separable. To solve this problem, we could apply the kernel method in the future.

Result Analysis

Performance metrics
As a multi-class classification problem, a natural way to evaluate the performance is the accuracy. As for the recall and precision, there are two ways to calculate them:
Macro averaged: calculate recall/precision for all classes individually and then average them.

Micro averaged: calculate class wise true positive and false negative and then use that to calculate overall recall.

Since the dataset is awfully unbalanced,the macro averaged will yield lower results. The reason is that each class will have the same weight. To address the unbalance problem, we choose to use the macro averaged ones. Furthermore, we choose to report the F1 score to balance the recall and precision.

Random forest

We applied random forest model with three variations, the performances of all classifiers are as followed:

Model	Accuracy	F1
Random forest	61.89%	27.98%
Random forest with standardization	61.19%	27.73%
Random forest with oversampling	36.48%	15.82%

One improvement we tried is to use grid search to set the parameters of the random forest. By finding the parameters minimizing the training error, each tree's max depth, the feature used and the minimum leaf sizes are determined. The accuracy improvement on the training set is approximately 7% but on the testing set the accuracy decreased by 2%. This suggests that the grid search imposes an overfit problem to the dataset.

KNN

Model	Accuracy	F1
KNN	52.80%	23.80%
KNN with standardization	58.39%	27.11%
KNN with oversampling	12.51%	3.71%

Logistic Regression

Model	Accuracy	F1
Logistic Regression	58.39%	22.84%

Logistic Regression with standardization	59.09%	26.28%
Logistic Regression with oversampling	36.13%	14.23%

In the logistic regression model, we try to give higher penalties for the minority class. Grid search is applied to find the optimum number in the training set. The accuracy improved by around 2% while the F1 increased by 3%. However, the optimal weight does not deviate too much from the regular loss function.
We also tried to define the dependent variable quality as an ordinal variable and use the logistic regression to predict. The performance does not change too much.

Support Vector Machine

Model	Accuracy	F1
SVM	53.15%	19.05%
SVM with standardization	61.89%	26.30%
SVM with oversampling	42.26%	9.90%

In general, standardization boosts the performance of KNN and SVM. These two models are affected by the scale of the dataset so standardization can lead to more robust results.

Surprisingly, the SMOTE technique reduces the performance by a lot. By oversampling on the minority cases, the classifier tends to predict a lot of normal quality wrong. While the accuracy in the minority increased a little, the overall accuracy and F1 score still decreased. This suggests that the cost of misclassified majority samples of oversampling overweights the gain. This implies there may be some outlier of the quality score since it is fully subjective. Therefore, we do not recommend using oversampling.

Random forest and logistic regression yields the highest results among all cases. The decent performance of random forest may be its robustness in bagging. As for the logistic regression, since we can explicitly adjust the weight for unbalanced classes, the model can perform better on unbalanced classifications.

Implications

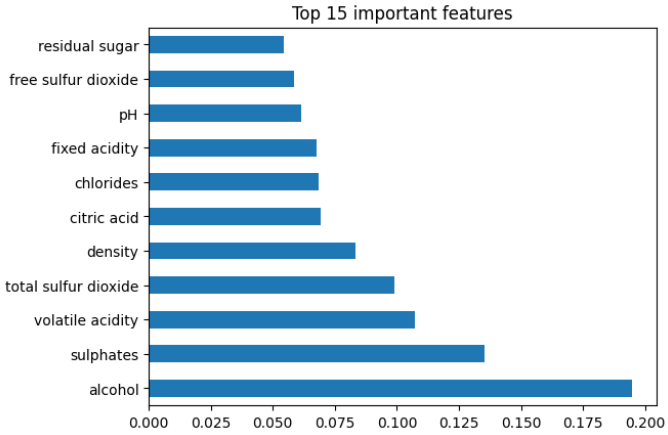
While it is important to get quantitative prediction result, it is also important to interpret the meaning of these models. To get explainable results, we analyzed the result from PCA, random forest and logistic regressions.

PCA

The PCA does not apply well in this dataset. The variance explained by the first factor is only around 30% and the weight of each variables hardly gives us any useful information.

Random forest

We calculated the feature importance based on mean decrease in impurity. We found that Alcohol, sulphates and volatile acidity are the three most important features. This indicates that these three factors are most likely to influence the quality of red wine.



Logistic Regression

We obtained the coefficients from the logistic regression, which can help us examine the impact of different variables. Firstly, since the data has already been standardized, the coefficients suggest the magnitude of the impact. pH, fixed acidity and alcohol seems to have the largest impact on the quality. Judging the sign of the coefficient, we can also know that alcohol and sulphates have a positive impact on quality. Also, fixed acidity and pH decreases the quality.

	coef	std err	t	P>|t|
const	-5.6570	0.024	-237.784	0.000
fixed acidity	-0.1122	0.066	-1.691	0.091
volatile acidity	0.0456	0.032	1.436	0.151
citric acid	0.0324	0.043	0.759	0.448
residual sugar	-0.0046	0.031	-0.147	0.883
chlorides	-0.0500	0.030	-1.696	0.090
free sulfur dioxide	0.0154	0.033	0.469	0.639
total sulfur dioxide	-0.0244	0.035	-0.707	0.480
density	0.0061	0.061	0.099	0.921
pH	-0.1097	0.044	-2.503	0.012
sulphates	0.0146	0.029	0.510	0.610
alcohol	0.0493	0.042	1.161	0.246

Discussion

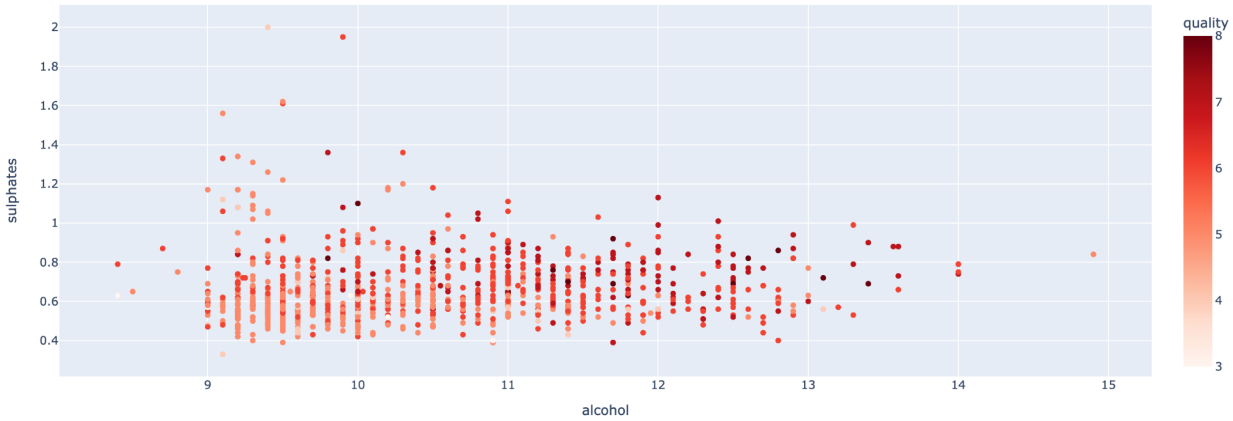
Based on the discussions above, we can locate the factors that have an impact on quality. After conducting research in the manufacturing process, we find that there is a trade off in every factor.

1. Sulphates is added as a preservative, and may cause headaches.Wines with lower volatile acidity need more sulfites than higher acidity wines. At pH3.6 and above, wines are much less stable, and sulfites are necessary for shelf-life.

2. Alcohol in red wine is usually between 9% and 14%, with an average of 11%. Too much alcohol is not possible for red wine.

The above two rules tell us there cannot be an ideal case where the quality is maximized. There is bound for alcohol and sulphates. Increasing fixed acidity would decrease pH,and reducing acidity needs more sulphates and sulfur dioxide. All of the pairs have a opposite effect on quality. There also may be other factors like cost, types of grapes being used that can constraint the improvement.

However, looking at the scatter plot below, red wine with 12-14 alcohol level and 0.8-1 sulphates seem to have higher quality than the rest. Our recommendation is then to control the two factors in this level which will make it more likely to gain high quality.



Limitations and future research

The biggest limitation of the problem is that the quality is purely subjective. Certain consumers may love/hate some kind of wine, which creates some outliers. It is hard to use mathematical techniques to distinguish them which makes the dataset noisy. Therefore, future research should focus on the collection of data and try to incorporate more dimensions. More data would also benefit the prediction performance of the model.

Reference

[1] https://www.kaggle.com/code/nkitgupta/evaluation-metrics-for-multi-class-classification
[2] https://winefolly.com/deep-dive/sulfites-in-wine/

Appendix:

This analysis includes a wide range of physicochemical properties of wine:

Fixed Acidity: This refers to the concentration of acids in the wine, mainly nonvolatile or fixed acids that don't evaporate easily.

Volatile Acidity: This measures the quantity of acetic acid in the wine. If the levels are too high, it can give the wine an unpleasant vinegar-like taste.

Citric Acid: Although present in small amounts, citric acid can impart a sense of 'freshness' and enhance the wine's flavor.

Residual Sugar: This is the quantity of sugar left after the fermentation process. Wines usually have at least 1 gram per liter, and those with more than 45 grams per liter are generally considered sweet.

Chlorides: This represents the level of salt in the wine.

Free Sulfur Dioxide: This is the free form of SO2 that exists in equilibrium between molecular SO2 and the bisulfite ion. It helps to prevent microbial growth and oxidation.

Total Sulfur Dioxide: This is the sum of the free and bound forms of SO2. At low concentrations, SO2 is mostly undetectable in wine, but when free SO2 concentrations exceed 50 ppm, SO2 becomes noticeable in the wine's aroma and taste.

Density: This measures how dense the wine is. The density of wine is similar to that of water, with variations depending on the percentage of alcohol and sugar content.

pH: This indicates the level of acidity or basicity of the wine on a scale from 0 (very acidic) to 14 (very basic), with most wines falling in the 3-4 range on the pH scale.

Sulphates: These are wine additives that can increase sulfur dioxide gas (SO2) levels, acting as an antimicrobial and antioxidant.

Alcohol: This is the percentage of alcohol content in the wine.