Predicting Wine Quality with Machine Learning

Group 9

Victoria Liu (vl292), Yanfei Dai (yd429), Yihong He (yh827), Yueyao Guo (yg588),
Haoqi Liu (hl2453)

Abstract

Our paper focuses on predicting wine quality using machine learning models in replacement of the traditional wine assessment method by using a combination of physicochemical and sensory tests, and the latter can only be performed by experts of wine tasting. In summary, we tested out four machine learning models: logistic regression, random forest, Support Vector Machine (SVM), and XGBoost. We trained and tested all of our models on a dataset containing information of wine samples from Portugal. The results of our project may be able to help the wine industry on aspects of sales and safety.

Introduction

Wine is becoming an increasingly popular drink, and along with people's growing interest in the wine industry, comes higher focuses and expectations on wine classification, especially the quality assessment of wine. Wine classification is not only used for pricing purposes, but also used for public safety purposes due to creation of low quality, unsafe, and usually illegal wine. Two of the most important components of wine quality assessment are physicochemical and sensory tests. Sensory tests are only to be performed by humans (wine tasting experts), and the relationship between sensory and physicochemical tests are unclear, leading to the purpose of our study.

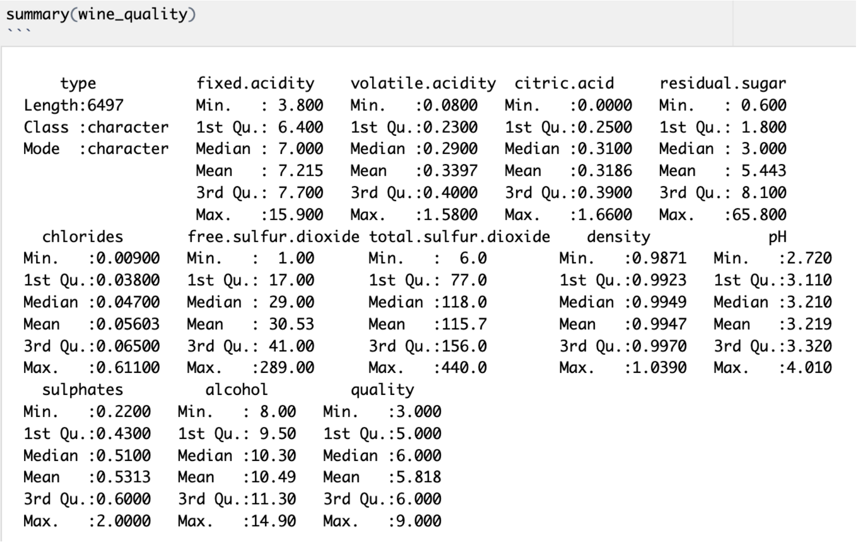
Due to the large amounts of effort needed in order to access the quality of wine in contrast to its high demands, our team wanted to use technology to make the wine classification process easier. For our study, we used a dataset, containing a long list of extensive information about red and white vinho verde wine samples, collected by Paulo Cortez and his research group. The dataset describes quality as a number from 0 to 10, with 0 being the worst quality and 10 being the best quality. We first tested out multiple machine learning models to predict the quality of wine, using the other 12 features included in the dataset. After the initial modeling, we also tuned the parameters of each of these models in hopes of getting a higher prediction accuracy.

The next few sections of our paper discuss our methodologies, results, and conclusions in more detail.

Machine Learning Methods

Data Distribution and Structure:

- The dataset has 6497 observations.
- Quality ranges from 3 to 9, and scores are from 0 to 10. The median score is 6, indicating a central tendency towards medium quality wines.
- Alcohol content varies widely from 8% to 14.9%, with the median at 10.3%.
- Acidity levels (fixed, volatile, and citric) and sugar levels vary significantly, which is typical for such data, reflecting different styles of wine.
- Sulfur dioxide levels also show significant variation, which impacts the preservation and aging processes of wines.
- Density and pH are key factors that vary less but are crucial for wine quality.



~~~
summary(wine_quality)
```

     type           fixed.acidity    volatile.acidity  citric.acid     residual.sugar
 Length:6497        Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
 Class :character   1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
 Mode  :character   Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
                    Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
                    3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
                    Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density             pH
 Min.   :0.00900   Min.   :  1.00      Min.   :  6.0       Min.   :0.9871   Min.   :2.720
 1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0       1st Qu.:0.9923   1st Qu.:3.110
 Median :0.04700   Median : 29.00      Median :118.0       Median :0.9949   Median :3.210
 Mean   :0.05603   Mean   : 30.53      Mean   :115.7       Mean   :0.9947   Mean   :3.219
 3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0       3rd Qu.:0.9970   3rd Qu.:3.320
 Max.   :0.61100   Max.   :289.00      Max.   :440.0       Max.   :1.0390   Max.   :4.010
   sulphates         alcohol         quality
 Min.   :0.2200   Min.   : 8.00   Min.   :3.000
 1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000
 Median :0.5100   Median :10.30   Median :6.000
 Mean   :0.5313   Mean   :10.49   Mean   :5.818
 3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
 Max.   :2.0000   Max.   :14.90   Max.   :9.000
~~~

Given that the target variable, quality, is a numerical score, we plan to try both regression models and classification models. For regression models, we will use random forest regressor which is more robust than linear regression when handling non-linear relationships, and also XGBoost which is similar to random forest but often provides better predictive accuracy through boosting technique. For classification models, we will use logistic regression as a starting point.

It can provide a baseline for performance and is useful for understanding the influence of several independent variables on a single outcome variable. Then, we decided to use SVM, which can be tuned for ordinal regression and might provide good margins of separation between quality classes.

Data Preprocessing:

First of all, we convert the "quality" column into a categorical variable to simplify the analysis. Quality scores are converted to a categorical factor, this conversion aligns with the practical understanding of quality as a categorical, rather than continuous. Then we use one-hot encoding and scaling to deal with specific features. For example, we changed the “type” into a binary numeric format 0/1 that reflects red/white for the wine type.

Feature Engineering:

We create interaction terms to provide the model with insights about the combined effects of two or more variables by multiplying them. For example, interactions between total sulfur dioxide and free sulfur dioxide are introduced into the dataset. This interaction term can help to model the non-additive effects of free and total sulfur dioxide on wine quality. We believe that even if both high total and low free sulfur dioxide levels are generally unfavorable, their combined effect might differ depending on their proportions. We also introduce the interaction effect between Alcohol and Volatile Acidity. The rationale behind this is that Volatile acidity at high levels can lead to an unpleasant vinegar taste, while alcohol contributes to the body, flavor, and preservation of wine. The balance between these factors is crucial for sensory quality.

Furthermore, we included polynomial terms into our dataset to help capture non-linear relationships between features and target variables. We introduced polynomial terms on Fixed Acidity and Citric Acid. This can enhance model performance significantly when the relationship is not linear.

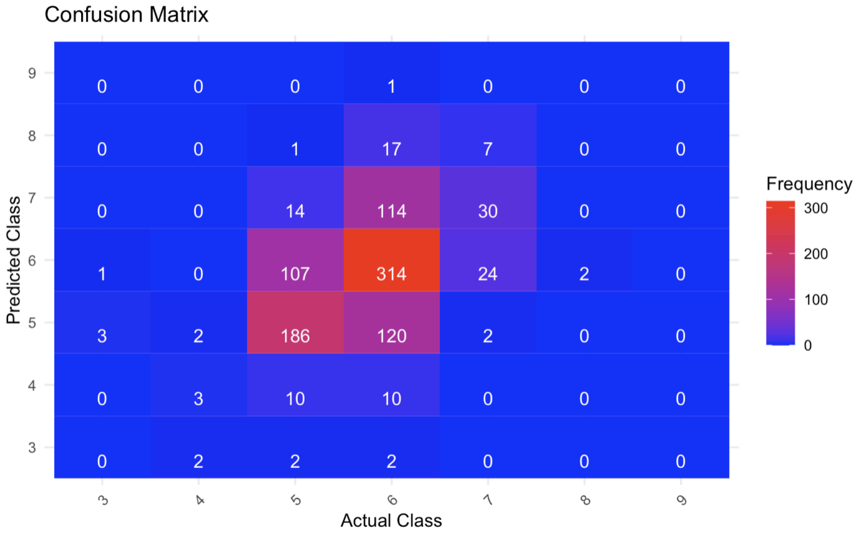
Model Evaluations:

1. Logistic regression

The first model we use for this project is Logistic Regression, which is employed to model the probability distribution of the wine quality classes. The model is built using the multinom function from the nnet package, which handles multinomial logistic regression, suitable for categorical outcome variables like wine quality.

The hyperparameter we choose to tune for this model is decay, which represents regularization strength tuned to help prevent overfitting and improve model generalization. This is achieved through a grid search over multiple decay values: 0.01, 0.05, 0.1, 0.5.

Tuning the decay parameter helps in managing the complexity of the model and balancing between bias and variance.



Confusion Matrix

Predicted Class \ Actual Class	3	4	5	6	7	8	9
9	0	0	0	1	0	0	0
8	0	0	1	17	7	0	0
7	0	0	14	114	30	0	0
6	1	0	107	314	24	2	0
5	3	2	186	120	2	0	0
4	0	3	10	10	0	0	0
3	0	2	2	2	0	0	0

2. Random Forest

We picked the Random Forest model to perform the classification task related to predicting wine quality. The Random Forest model operates by creating multiple decision trees during training and outputs the mode of the classifications. We chose to use random forest to predict wine quality for its robustness against overfitting, especially in scenarios with complex data structures like wine quality attributes.

We tuned the hyperparameter mtry for this model, which is the number of variables considered at each split in a tree. We believe tuning this hyperparameter can optimize model accuracy and prevent overfitting. Adjusting mtry allows the model to explore different combinations of features at each split, potentially discovering more informative splits based on feature interactions.

3. Support Vector Machine (SVM)

SVM is used with a radial basis function (RBF) kernel to handle non-linear relationships in the data. It categorizes the wine qualities by constructing a hyperplane in a high-dimensional space.

Two critical parameters, cost and gamma, are tuned. Cost controls the penalty of the error term, and gamma defines the influence of a single training example. These are tuned over a range of values for cost from 5 to 15, increment by 2, and gamma 0.75,1,1.25.

The tuning of cost and gamma is crucial for the SVM's ability to manage the trade-off between bias and variance and to define how non-linear the decision boundary should be, which is vital for achieving good generalization on unseen data.

4. Xgboost

We choose to use XGBoost because of its efficiency and effectiveness at scale. It uses a gradient boosting framework and is configured to solve a multi-class classification problem.

Several parameters are tuned, including eta, the learning rate, max_depth, the maximum depth that a tree can reach, subsample, and colsample_bytree, which control the fraction of the dataset and features used per tree.

These parameters are tuned to control the training process's speed and effectiveness, helping to prevent overfitting while ensuring that the model is adequately complex to capture the underlying patterns in the data.

Results

The performance of the four machine learning models—Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost—was evaluated based on their accuracy in predicting wine quality as a factor. The accuracy was assessed at two tolerance levels: Tolerance 0, which only considers correct predictions, and Tolerance 1, which allows a discrepancy of one quality level for all the wine quality predictions. The results were also visualized in a bar graph comparing model accuracies across different tolerance levels.

Logistic Regression

Tolerance 0: 54.72%.

Tolerance 1: 94.66%

Random Forest

Tolerance 0: 67.66%

Tolerance 1: 96.61%

Support Vector Machine (SVM)

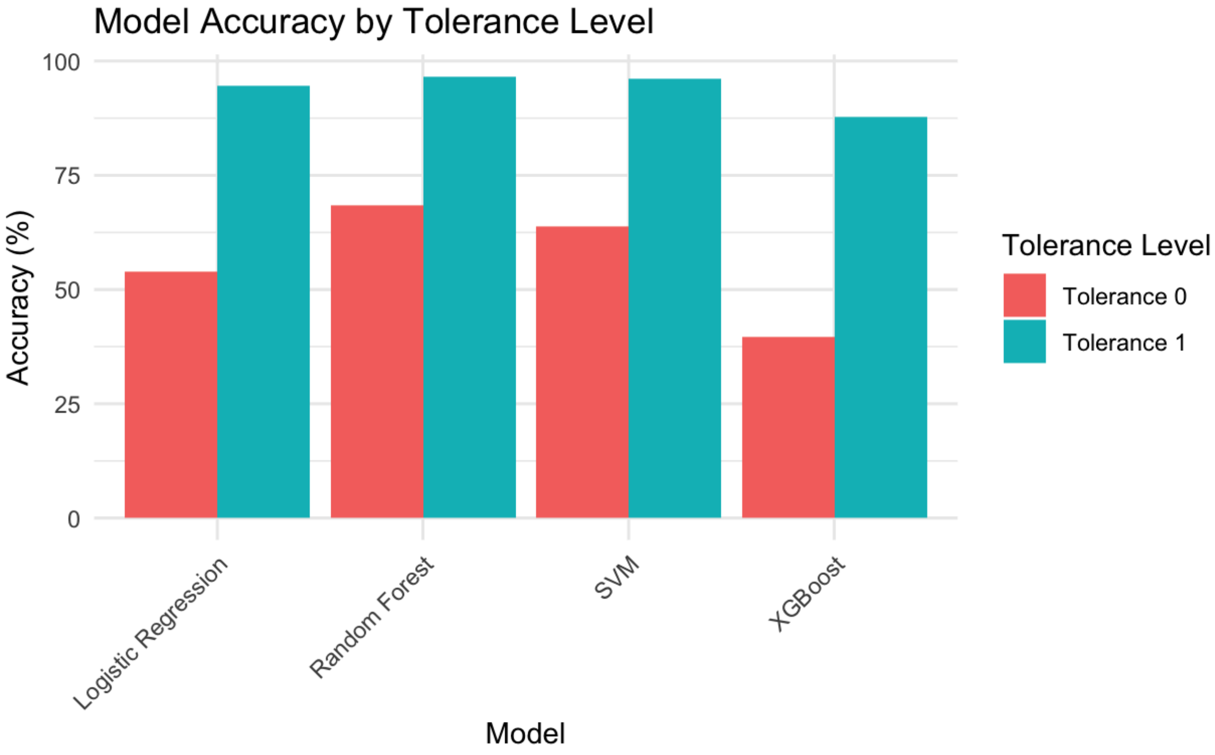
Tolerance 0: 63.76%

Tolerance 1: 96.20%

XGBoost

Tolerance 0: 38.47%

Tolerance 1: 87.25%



Conclusion

To conclude, in this machine learning prediction model, random forest has highest accuracy with 67.66% accuracy for exact matches (Tolerance 0) and 96.61% when allowing for a discrepancy of one quality level (Tolerance 1). We believe that the advantage of random forest in this ML is its ability to manage the more complex nonlinear relationships in the data without over-fitting. This indicates that the model is particularly suitable for predicting some data sets with inherent complexity and interactivity. Secondly, the Accuracy of the SVM model is also high. Although the accuracy of this model is 63.76%, and under tolerance 1 is 96.20%.

According to this model's Data Manipulation process, we found that it is necessary to adjust its cost and database gamma parameters in order to define the decision boundary. This means that SVM is accurate under optimal hyperparameters. On the other hand, the accuracy of the prediction of the logistic regression model is 54.72% when the model's tolerance is 0 and 94.66% when tolerance is 1. The accuracy level of this model is substantially less effective than the previous two ones, and the model does not seem as robust against overfitting as integrated methods such as random forests. However, this discrepancy still demonstrates the relevance of this model in offering a foundational understanding of data in ML processes. Finally, the performance was lower at tolerance 0, with an accuracy of 38.47%, but XGBoost improved to 87.25% at tolerance 1. This shows that while XGBoost may be difficult to classify precisely in this case, it can still provide useful approximations. The adjustment of parameters such as eta, maximum depth, and subsample size has a good effect on balancing the trade-off between bias and variance. While random forests are the most efficient model overall, each machine learning model has advantages. The study therefore also highlights the importance of model tuning and selecting appropriate machine learning algorithms based on the specific characteristics of the data and the accuracy required for the task.

In general, based on this machine learning, we believe that ML can digitize the evaluation of wine quality and change the traditional evaluation process based on subjective evaluation by expert tasting. At the same time, these technologies can improve not only the accuracy of the quality assessment but also simplify the production and quality control processes, which can lead to the reduction of the costs and the increase of consumer trust.

R Code:

https://github.com/yihonghhe/stsci5740FinalProject