Predicting Wine Quality by Statistical Machine Learning

[Abstract] *In the essay, a series of models are proposed to predict wine quality, providing support for wine expert and advice for customers. Efficient and comprehensive as possible, accurate result will help wine society. The essay first utilizes non-parametric models to provide an intuitive classification method with excellent results from cutting-edge techniques. Next the essay carries parametric part by 1) a polynomial model with important product item; 2) utilize natural cubic spline and smoothing spline to check every variable. The essay finally makes comparison between red wine and white wine characteristics, discovering significant differences between the two differing kinds.*

1. Data Exploration

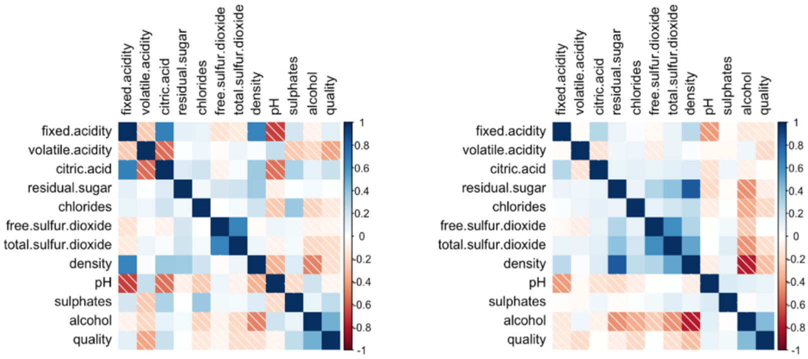


Figure 1 Red Wine Correlation Matrix (left), White Wine Correlation Matrix (right)

The main target of the essay is to provide excellent model to predict the quality of wine and seek into the wine characteristics.

The essay has two different datasets, i.e. red wine and white wine (Cortez et al., 2009). As Figure 1, both correlation matrices have similar correlation tendency, but neither is too significant. Both datasets confront non-uniform distribution problem. Only 0.5% of data is Class 3 (the lowest quality), and nearly no wine data hits the highest class that only happens in white wine, as Figure 2. Thus, we divide the wine into Good wine and Bad wine, in which have Class 3, 4, 5 and Class 6, 7, 8, 9 respectively, by which the target value is now more uniform for the robustness of further modelling.

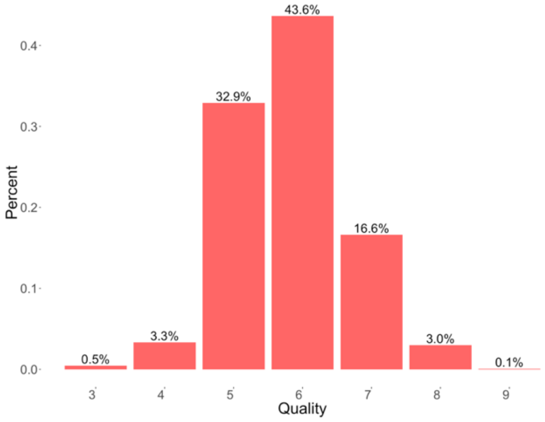


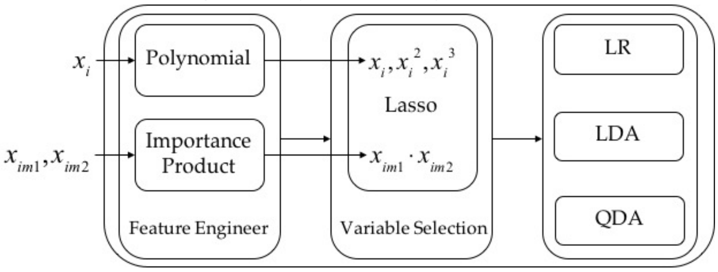
Figure 2 Quality Distribution

2. Model Solution

2.2 Parametric approach

2.2.1 Polynomial Classification

Figure 3 Polynomial Classification Framework



In polynomial models, the essay extends attributes by adding polynomial of degree up to three for each attribute (Keating & Cherry., 2011). The essay also picks the most two influencing attributes, making their product as a new predictor, by visualizing the importance of each attributes using random forest in previous work (Jaccard, 2003). For selection methods, best subset selection does not fit our massive data amount; therefore, the paper chooses Lasso to determine the best suitable model.

Table 1 Polynomial Classification

	LR	LDA	QDA
red wine	0.77	0.77	0.78
white wine	0.78	0.79	0.78

Table 2 Spline Classification

	Natural Spline	Smoothing Spline
red wine	0.77	0.77
white wine	0.77	0.79

Most of the parametric classification approaches work well and have little differences between red and white as can be showed by Table 1. Since the paper add polynomial items and product variable into models, the distribution is not strict normal, so that it cannot expect an even better result by them.

2.2.2 Spline Classification

Spline classification goes from the original predictors. We set natural cubic spline transform and smoothing spline transform to all variables respectively to acquire different classification result. The results are showed in Table 2. Same as previous work, wine quality does have high degree correlations with predictors, and smoothing spline with penalty on twice differentiable functions concerns more on fluctuations and thus lead to better result with white wine, who holds more data to fit non-linear part.

2.1 Non-parametric approach

Parametric approach is not as good as non-parametric method, where the paper proposes seven models and get the accuracy results as following table shows.

Table 3 Non-parametric Approach

	Tree	Bagging	Random Forest	Boosting	SVM	Neural Network	**XGBoosting**
Red Wine	0.76	0.84	0.85	0.81	0.81	0.78	**0.86**
White Wine	0.73	0.84	0.86	0.84	0.81	0.77	**0.85**

Both datasets can be explained well in these models, among which Random Forest and XGBoosting keep the relative advantage among all the models. Bagging and Boosting also show good ability to classify the wine, in which Boosting performing better may result from larger white wine set that strengthens boosting. Random Forest is only slightly better than Bagging, due to the fact that these two methods have almost the same intuition, but the former provides random selection in each split, reducing overfitting. Support vector machine performance is also fairly good under radial basis kernel, a similarity measure. It concerns interaction between observations.

Here the essay introduces two cutting-edge classifiers, i.e. Neural Network, XGBoosting. The essay hereby implements a 2-hidden-layer fully connected backpropagation neural network. In hidden layer, it has 5 and 3 neurons, respectively and uses sigmoid function as activation function, whose graph can be checked in appendix. Thus, the non-linear relationship among variables will be covered comprehensively (Cortez et al., 2009). Nonetheless, it performs only a little bit better than simple tree. It is normal when data is insufficient for neural network, where effects between response and predictors cannot be fully explained.

The XGBoosting (Chen & Guestrin, 2016) has strong improvements towards Boosting, which boosts trees with subjectively chosen depth. Boosting is trying to gather the functions among all trees, which are acquired by fitting the residuals from previous tree. Intuitively, a tree should move only a small path from its predecessor, so the whole tree set can get the optimum appropriately at last. However, Subjective depth without complexity concern may prevent from getting global optimum. XGBoosting can, in contrast, decide the depth by minimizing a target function below. It concerns not only loss indicating how model fits training data, but also model complexity including the number of leaves and L2 term of residuals.

$$Obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i \right) + \sum_{K=1}^{K} \left(\gamma T + \frac{1}{2} \lambda \|w\|^2 \right)$$

$l(y_i, \hat{y}_i)$ is the fitting loss function, T is the number of leaves and $\|w\|^2$ indicates the L2 term of residuals, whereas γ and λ are corresponding penalty coefficients. From that, boosting depth is controlled with respect to both fitting and complexity that reduces overfitting as possible.

The essay tests other XGBoosting result, i.e. 1) precision: 0.85 for how accurate out of those predicted positive; 2) recall 0.87 for how accurate out of real positive data, and 3) F1: 0.86 for combining both of them. All of these show robustness and incredible prediction power of XGBoosting. The ROC curve is presented in Appendix, where it is easy to find the model is even approaching the top.

3. Comparison Analysis

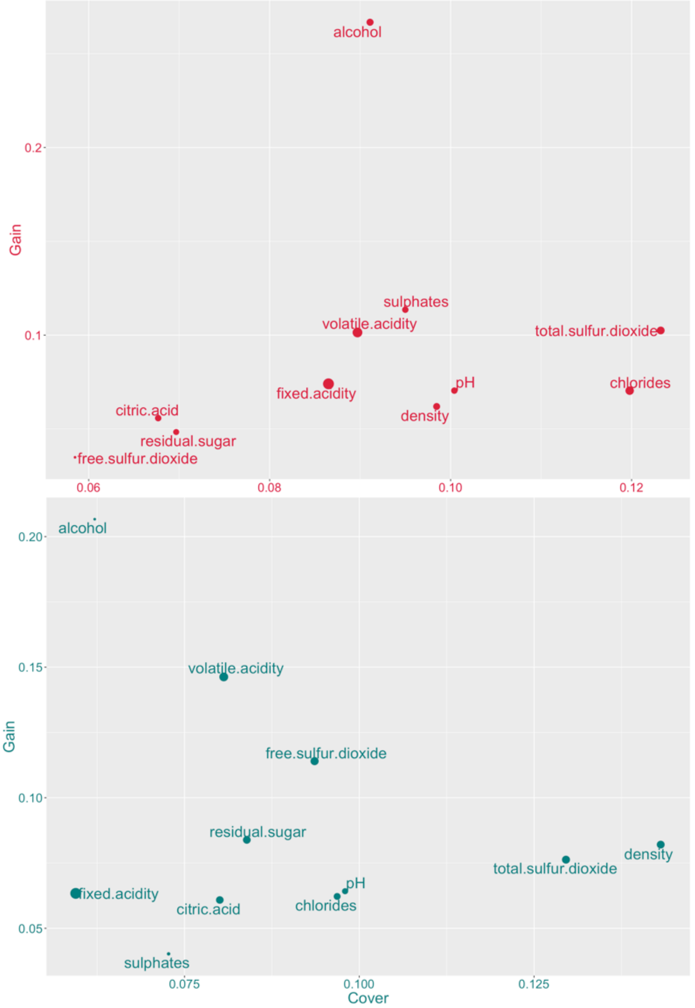
The essay also provides analysis into wine characteristics. By empirical logistic regression comparison to all original variables (see below in Table 4), both red wine and white wine take emphasis on *alcohol*, *sulphates, volatile acidity* and *free.sulfur.dioxide*. The first influences white more negatively, while other three more positively affect red wine. The *total.sulfur.dioxide* takes an significant negative impact only on red, as well as *citric acid* and *chlorides*, with less significance, while *residual.sugar*, *density* and *pH* significantly influence white wine. Unlike *volatile.acidity*, *fixed.acidity* does not show any impact with both kinds.

Table 1 Wine Comparison

	Red Wine	*White Wine*
alcohol	**0.867** ***	**0.7429** ***
sulphates	**2.7951** ***	**1.797** **
volatile.acidity	**-3.282** ***	**-6.459** ***
free.sulfur.dioxide	**0.022** **	**0.010*****
total.sulfur.dioxide	**-0.017** ***	-0.001
citric.acid	**-1.274** *	0.116
chlorides	**-3.916** *	0.8852
residual.sugar	0.055	**0.170** ***
density	-50.932	**-270.9*****
pH	-0.381	**1.09***
fixed.acidity	0.136	0.036

*** (99.9%) ** (99%) * (95%) ' (90%)

Figure 4 Important Variables, Red Wine (Red Upper), White Wine (Green Lower)



Obviously, *alcohol* is essential to all wine, but differing kinds of wine have differing the second concern. For red wine it is *sulphates*, but *volatile.acidity* for white wine. Red wine focuses on the *sulfur.dioxide*s, whereas *sugar* and *density* may be vital concern of white wine, which implies we care more how sweet white wine is.

We also find similar result XGBoosting importance detection. In Figure 4 for points, the higher represents the more contribution to the model. (Gain means the contribution of the feature by its splits). The righter indicates the number of observations related to this feature. And a larger point means more nodes splits this feature. The *alcohol* which is higher contribute mostly for red and white wine. In upper figure, *total.sulfur.dioxide* and *chlorides* are right to the end, whereas in lower figure, *volatile.acidity*, which is more important to white wine showed in the table, dominates others except alcohol. Other orders are also presented above. It is interesting to see although white wine view *alcohol* as a very important predictor, it is hardly to see the times it splits the data, as the point is so small. It can even show that this variable is rather essential, with higher efficiency.

4. Conclusion

The essay focuses on binary classification, which is due to the small amount of data. The data is split by threshold 6. That is quality Larger than and equal to 6 is regarded as “Good”, “Bad” on the other side.

Non-parametric models with stronger techniques, i.e. XGBoosting overwhelm parametric models, which, in terms of accuracy particularly, is the most suitable model for both wines. On the other hand, it is easy to understand the result, especially for features importance comparison. Parametric models, on the contrary, does not provide as good result as the others. Nonetheless, coefficients of Logistic Regression, by which we compare issues of two differing wine, and describe characteristics of them, is still meaningful.

The essay hereby introduces XGBoosting as a supreme model for wine quality prediction. The essay recommends its use in wine recommendation or initially classification. Experts can refer to our result of wines or double check their subjective predictions.

As for wines quality, the essay believes degree of alcohol is vital to wine quality. The more sulphates will contribute to quality for both kinds, but mainly for red wine. The higher volatile acidity will lead to bad quality, but especially white wine. Free sulfur dioxide will enhance wines but the total should be controlled, red wine particularly. For better red wine, producer should also take care of citric acid and chlorides. And white wine tasters also care residual sugar, the density of wine and its pH value.

Due to the small amount of data, especially in the extreme cases. exceeding unbalance classes are severely influence model robustness if modeling multiclassification. It also prevents better result from Neural Network implemented before. Dealing with more data will be further wish to achieve even greater results. At the same time, wine recommendation towards individuals can be also interesting to explore.

[APPENDIX I: Visualization Graphs]

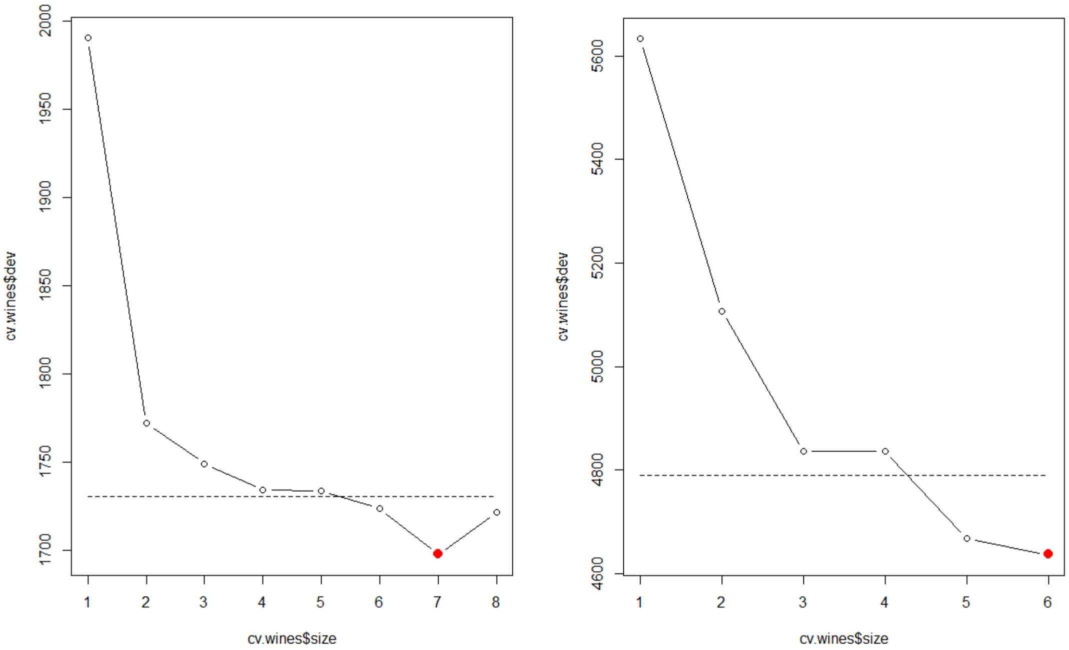


Figure 1 Cross-validation results for tree size, Red Wine (left), White Wine (right)

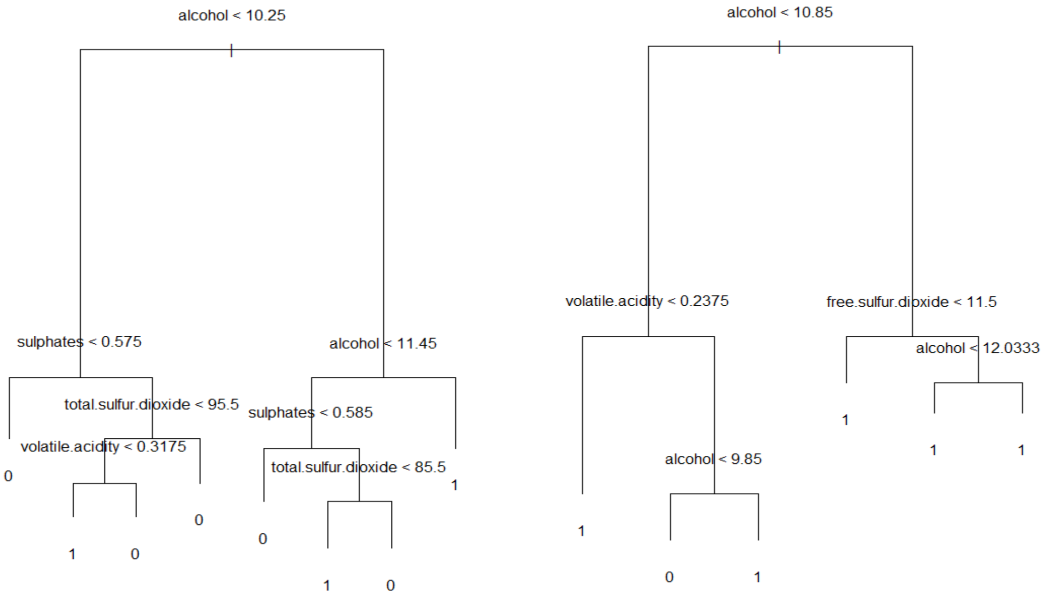

alcohol < 10.25
sulphates < 0.575
alcohol < 11.45
total.sulfur.dioxide < 95.5
volatile.acidity < 0.3175
sulphates < 0.585
total.sulfur.dioxide < 85.5
0
1
0
0
0
1
1
0
alcohol < 10.85
volatile.acidity < 0.2375
free.sulfur.dioxide < 11.5
alcohol < 12.0333
alcohol < 9.85
1
1
1
1
0
1

Figure 2 Pruned tree diagram, Red Wine (left), White Wine (right)

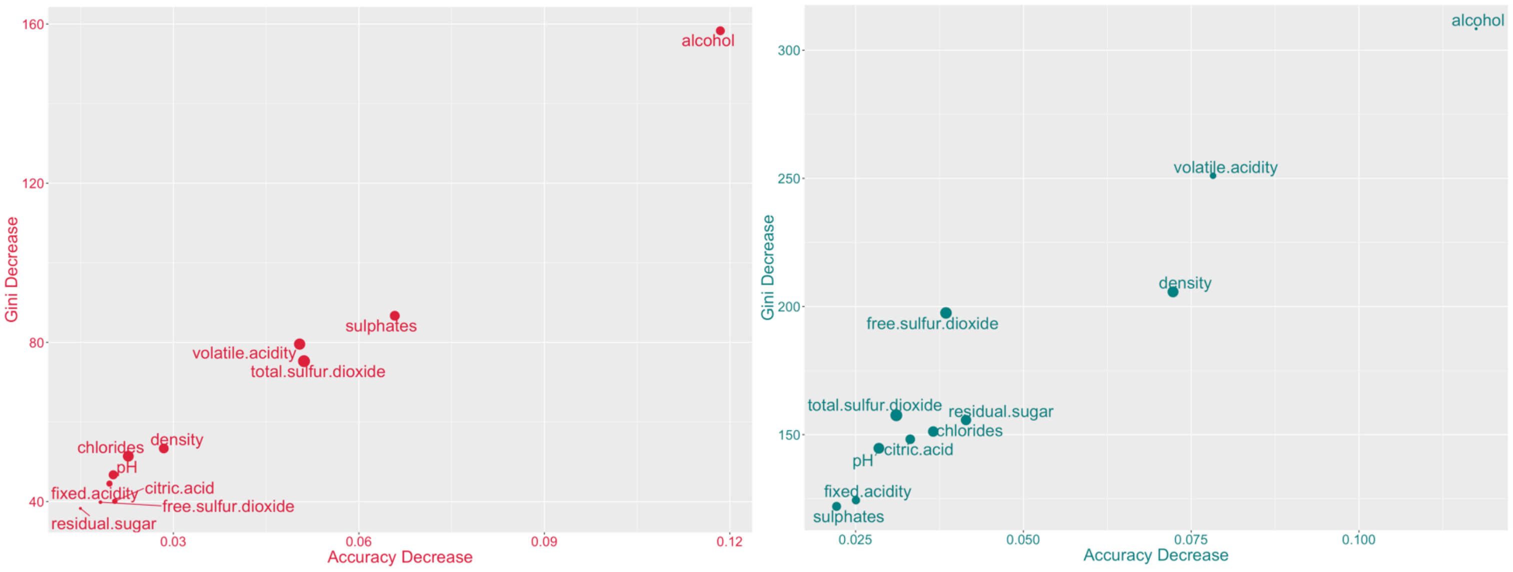


Figure 3 Important Variables by Random Forest, Red Wine (Red Left), White Wine (Green Right)

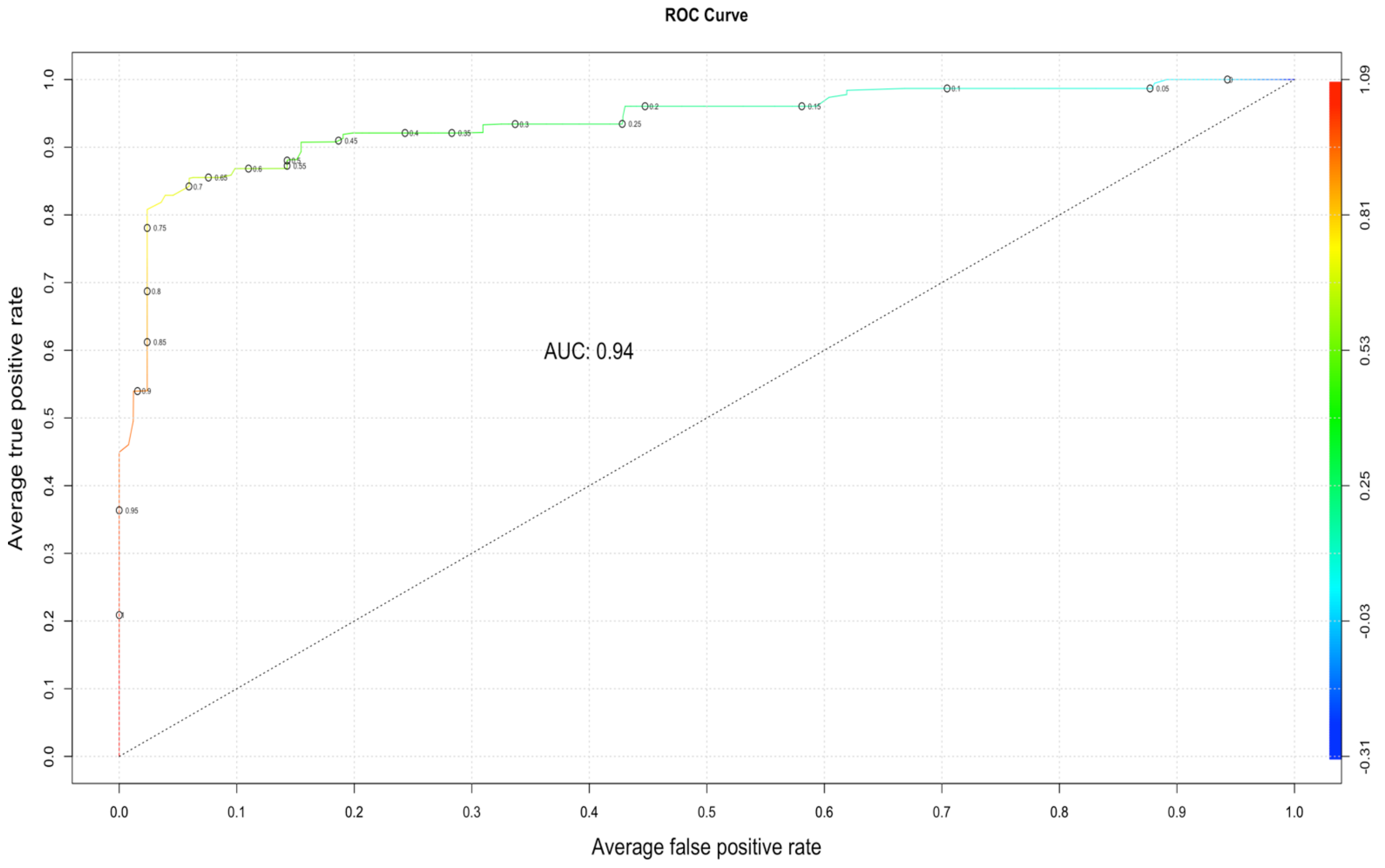


Figure 4 ROC for XGBoosting

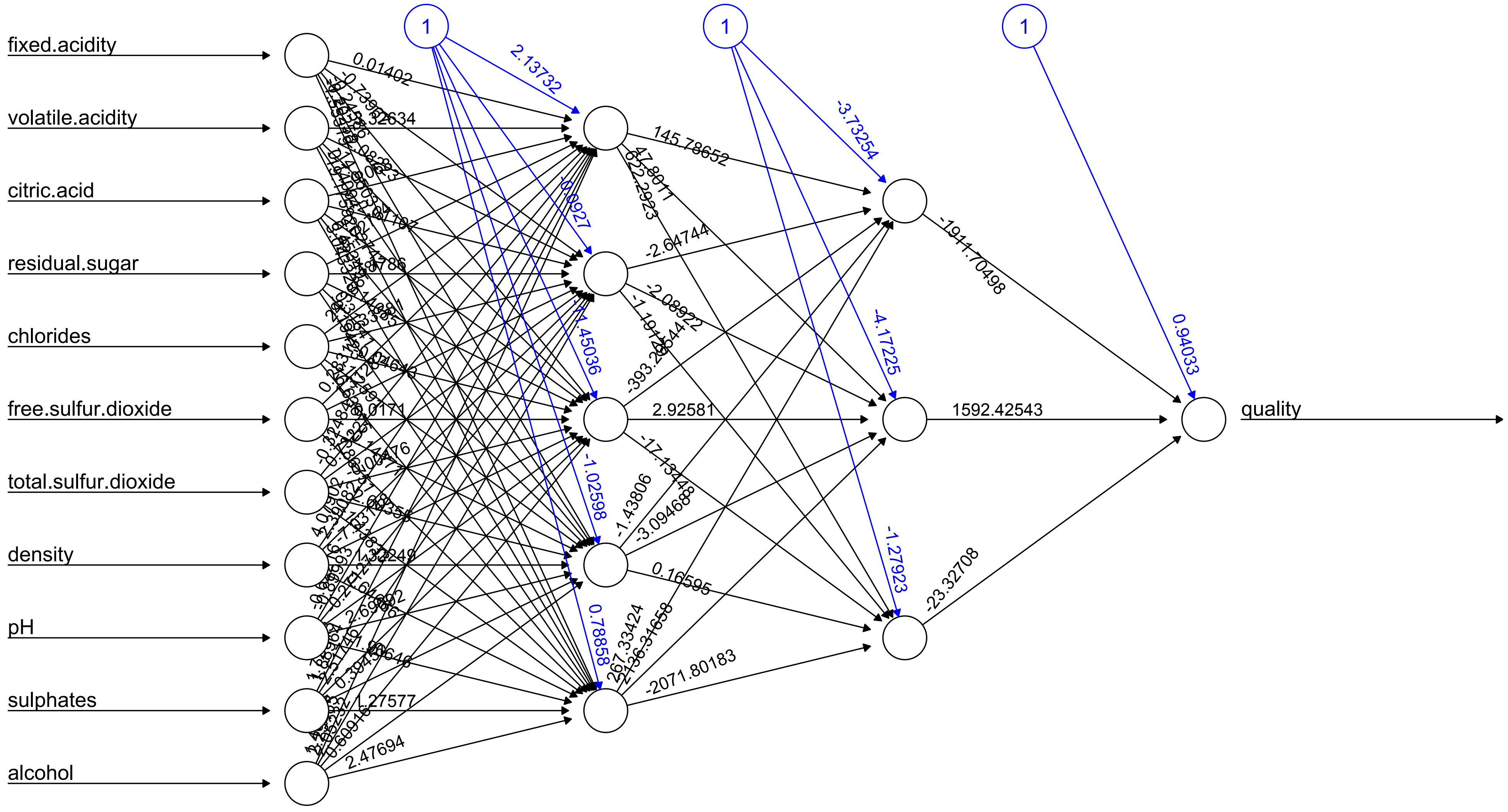


Figure 5 Neural Network Model

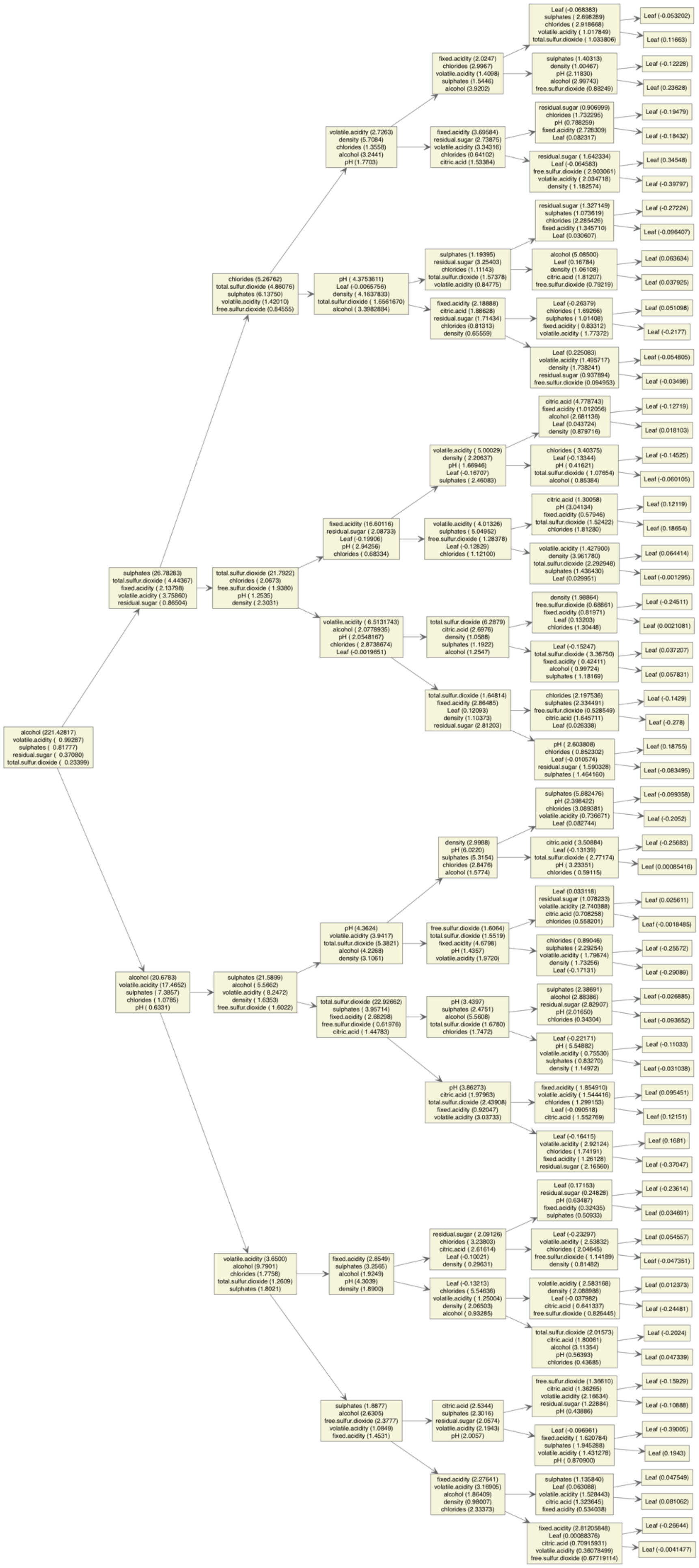

alcohol (221.42817) volatile.acidity (0.99287) sulphates (0.81777) residual.sugar (0.37080) total.sulfur.dioxide (0.23399)
sulphates (26.78283) total.sulfur.dioxide (4.44367) fixed.acidity (2.13798) volatile.acidity (3.75860) residual.sugar (0.86504)
chlorides (5.26762) total.sulfur.dioxide (4.86076) sulphates (6.13750) volatile.acidity (1.42010) free.sulfur.dioxide (0.84555)
volatile.acidity (2.7263) density (5.7084) chlorides (1.3558) alcohol (3.2441) pH (1.7703)
fixed.acidity (2.0247) chlorides (2.9967) volatile.acidity (1.4098) sulphates (1.5446) alcohol (3.9202)
Leaf (-0.068383) sulphates (2.698289) chlorides (2.918668) volatile.acidity (1.017849) total.sulfur.dioxide (1.033806)
Leaf (-0.053202)
Leaf (0.11663)
sulphates (1.40313) density (1.00467) pH (2.11830) alcohol (2.99743) free.sulfur.dioxide (0.88249)
Leaf (-0.12228)
Leaf (0.23628)
fixed.acidity (3.69584) residual.sugar (2.73875) volatile.acidity (3.34316) chlorides (0.64102) citric.acid (1.53384)
residual.sugar (0.906999) chlorides (1.732295) pH (0.788259) fixed.acidity (2.728309) Leaf (0.082317)
Leaf (-0.19479)
Leaf (-0.18432)
residual.sugar (1.642334) Leaf (-0.064583) free.sulfur.dioxide (2.903061) volatile.acidity (2.034718) density (1.182574)
Leaf (0.34548)
Leaf (-0.39797)
pH (4.3753611) Leaf (-0.0065756) density (4.1637833) total.sulfur.dioxide (1.6561670) alcohol (3.3982884)
sulphates (1.19395) residual.sugar (3.25403) chlorides (1.11143) total.sulfur.dioxide (1.57378) volatile.acidity (0.84775)
residual.sugar (1.327149) sulphates (1.073619) chlorides (2.285426) fixed.acidity (1.345710) Leaf (0.030607)
Leaf (-0.27224)
Leaf (-0.096407)
alcohol (5.08500) Leaf (0.16784) density (1.06108) citric.acid (1.81207) free.sulfur.dioxide (0.79219)
Leaf (0.063634)
Leaf (0.037925)
fixed.acidity (2.18888) citric.acid (1.88628) residual.sugar (1.71434) chlorides (0.81313) density (0.65559)
Leaf (-0.26379) chlorides (1.69266) sulphates (1.01408) fixed.acidity (0.83312) volatile.acidity (1.77372)
Leaf (0.051098)
Leaf (-0.2177)
Leaf (0.225083) volatile.acidity (1.495717) density (1.738241) residual.sugar (0.937894) free.sulfur.dioxide (0.094953)
Leaf (-0.054805)
Leaf (-0.03498)
total.sulfur.dioxide (21.7922) chlorides (2.0673) free.sulfur.dioxide (1.9380) pH (1.2535) density (2.3031)
fixed.acidity (16.60116) residual.sugar (2.08733) Leaf (-0.19906) pH (2.94256) chlorides (0.68334)
volatile.acidity (5.00029) density (2.20637) pH (1.66946) Leaf (-0.16707) sulphates (2.46083)
citric.acid (4.778743) fixed.acidity (1.012056) alcohol (2.681136) Leaf (0.043724) density (0.879716)
Leaf (-0.12719)
Leaf (0.018103)
chlorides (3.40375) Leaf (-0.13344) pH (0.41621) total.sulfur.dioxide (1.07654) alcohol (0.85384)
Leaf (-0.14525)
Leaf (-0.060105)
volatile.acidity (4.01326) sulphates (5.04952) free.sulfur.dioxide (1.28378) Leaf (-0.12829) chlorides (1.12100)
citric.acid (1.30058) pH (3.04134) fixed.acidity (0.57946) total.sulfur.dioxide (1.52422) chlorides (1.81280)
Leaf (0.12119)
Leaf (0.18654)
volatile.acidity (1.427900) density (3.961780) total.sulfur.dioxide (2.292948) sulphates (1.436430) Leaf (0.029951)
Leaf (0.064414)
Leaf (-0.001295)
volatile.acidity (6.5131743) alcohol (2.0778935) pH (2.0548167) chlorides (2.8738674) Leaf (-0.0019651)
total.sulfur.dioxide (6.2879) citric.acid (2.6976) density (1.0588) sulphates (1.1922) alcohol (1.2547)
density (1.98864) free.sulfur.dioxide (0.68861) fixed.acidity (0.81971) Leaf (0.13203) chlorides (1.30448)
Leaf (-0.24511)
Leaf (0.0021081)
Leaf (-0.15247) total.sulfur.dioxide (3.36750) fixed.acidity (0.42411) alcohol (0.99724) sulphates (1.18169)
Leaf (0.037207)
Leaf (0.057831)
total.sulfur.dioxide (1.64814) fixed.acidity (2.86485) Leaf (0.12093) density (1.10373) residual.sugar (2.81203)
chlorides (2.197536) sulphates (2.334491) free.sulfur.dioxide (0.528549) citric.acid (1.645711) Leaf (0.026338)
Leaf (-0.1429)
Leaf (-0.278)
pH (2.603808) chlorides (0.852302) Leaf (-0.010574) residual.sugar (1.590328) sulphates (1.464160)
Leaf (0.18755)
Leaf (-0.083495)
alcohol (20.6783) volatile.acidity (17.4652) sulphates (7.3857) chlorides (1.0785) pH (0.6331)
sulphates (21.5899) alcohol (5.5662) volatile.acidity (8.2472) density (1.6353) free.sulfur.dioxide (1.6022)
pH (4.3624) volatile.acidity (3.9417) total.sulfur.dioxide (5.3821) alcohol (4.2268) density (3.1061)
density (2.9988) pH (6.0220) sulphates (5.3154) chlorides (2.8476) alcohol (1.5774)
sulphates (5.882476) pH (2.398422) chlorides (3.089381) volatile.acidity (0.736671) Leaf (0.082744)
Leaf (-0.099358)
Leaf (-0.2052)
citric.acid (3.50884) Leaf (-0.13139) total.sulfur.dioxide (2.77174) pH (3.23351) chlorides (0.59115)
Leaf (-0.25683)
Leaf (0.00085416)
free.sulfur.dioxide (1.6064) total.sulfur.dioxide (1.5519) fixed.acidity (4.6798) pH (1.4357) volatile.acidity (1.9720)
Leaf (0.033118) residual.sugar (1.078233) volatile.acidity (2.740388) citric.acid (0.708258) chlorides (0.558201)
Leaf (0.025611)
Leaf (-0.0018485)
chlorides (0.89046) sulphates (2.29254) volatile.acidity (1.79674) density (1.73256) Leaf (-0.17131)
Leaf (-0.25572)
Leaf (-0.29089)
total.sulfur.dioxide (22.92662) sulphates (3.95714) fixed.acidity (2.68298) free.sulfur.dioxide (0.61976) citric.acid (1.44783)
pH (3.4397) sulphates (2.4751) alcohol (5.5608) total.sulfur.dioxide (1.6780) chlorides (1.7472)
sulphates (2.38691) alcohol (2.88386) residual.sugar (2.82907) pH (2.01650) chlorides (0.34304)
Leaf (-0.026885)
Leaf (-0.093652)
Leaf (-0.22171) pH (5.54882) volatile.acidity (0.75530) sulphates (0.83270) density (1.14972)
Leaf (-0.11033)
Leaf (-0.031038)
pH (3.86273) citric.acid (1.97963) total.sulfur.dioxide (2.43908) fixed.acidity (0.92047) volatile.acidity (3.03733)
fixed.acidity (1.854910) volatile.acidity (1.544416) chlorides (1.299153) Leaf (-0.090518) citric.acid (1.552769)
Leaf (0.095451)
Leaf (0.12151)
Leaf (-0.16415) volatile.acidity (2.92124) chlorides (1.74191) fixed.acidity (1.26128) residual.sugar (2.16560)
Leaf (0.1681)
Leaf (-0.37047)
volatile.acidity (3.6500) alcohol (9.7901) chlorides (1.7758) total.sulfur.dioxide (1.2609) sulphates (1.8021)
fixed.acidity (2.8549) sulphates (3.2565) alcohol (1.9249) pH (4.3039) density (1.8900)
residual.sugar (2.09126) chlorides (3.23803) citric.acid (2.61614) Leaf (-0.10021) density (0.29631)
Leaf (0.17153) residual.sugar (0.24828) pH (0.63487) fixed.acidity (0.32435) sulphates (0.50933)
Leaf (-0.23614)
Leaf (0.034691)
Leaf (-0.23297) volatile.acidity (2.53832) chlorides (2.04645) free.sulfur.dioxide (1.14189) density (0.81482)
Leaf (0.054557)
Leaf (-0.047351)
Leaf (-0.13213) chlorides (5.54636) volatile.acidity (1.25004) density (2.06503) alcohol (0.93285)
volatile.acidity (2.583168) density (2.088988) Leaf (-0.037982) citric.acid (0.641337) free.sulfur.dioxide (0.826445)
Leaf (0.012373)
Leaf (-0.24481)
total.sulfur.dioxide (2.01573) citric.acid (1.80061) alcohol (3.11354) pH (0.56393) chlorides (0.43685)
Leaf (-0.2024)
Leaf (0.047339)
sulphates (1.8877) alcohol (2.6305) free.sulfur.dioxide (2.3777) volatile.acidity (1.0849) fixed.acidity (1.4531)
citric.acid (2.5344) sulphates (2.3016) residual.sugar (2.0574) volatile.acidity (2.1943) pH (2.0057)
free.sulfur.dioxide (1.36610) citric.acid (1.36265) volatile.acidity (2.16634) residual.sugar (1.22884) pH (0.43886)
Leaf (-0.15929)
Leaf (-0.10888)
Leaf (-0.096961) fixed.acidity (1.620784) sulphates (1.945288) volatile.acidity (1.431278) pH (0.870900)
Leaf (-0.39005)
Leaf (0.1943)
fixed.acidity (2.27641) volatile.acidity (3.16905) alcohol (1.86409) density (0.98007) chlorides (2.33373)
sulphates (1.135840) Leaf (0.063088) volatile.acidity (1.528443) citric.acid (1.323645) fixed.acidity (0.534038)
Leaf (0.047549)
Leaf (0.081062)
fixed.acidity (2.81205848) Leaf (0.00088376) citric.acid (0.70915931) volatile.acidity (0.36078499) free.sulfur.dioxide (0.67719114)
Leaf (-0.26644)
Leaf (-0.0041477)

Figure 6 XGBoosting Model

[APPENDIX II: Formula Presentation]

$$
\begin{aligned}
quality_{red} = {} & \beta_0 + \beta_1 fixed.acidity + \beta_2 volatile.acidity + \beta_3 free.sulfur.dioxide + \beta_4 total.sulfur.dioxide \\
& + \beta_5 alcohol + \beta_6 residual.sugar^2 + \beta_7 chlorides^2 + \beta_8 pH^2 + \beta_9 alcohol \cdot sulphates + \varepsilon
\end{aligned}
$$

$$
\begin{aligned}
quality_{white} = {} & \beta_0 + \beta_1 volatile.acidity + \beta_2 residual.sugar + \beta_3 chlorides + \beta_4 free.sulfur.dioxide + \beta_5 pH \\
& + \beta_6 sulphates + \beta_7 alcohol + \beta_8 fixed.acidity^2 + \beta_9 total.sulfur.dioxide^2 + \beta_{10} pH^2 + \varepsilon
\end{aligned}
$$

Fomula 1, Polynomial Classification Logistic Regression Model
quality here resembles its sigmoid form
(Red Wine Upper, White Wine Lower)

$$
\begin{aligned}
quality = {} & \beta_0 + \beta_1 fixed.acidity + \beta_2 volatile.acidity + \beta_3 citric.acid + \beta_4 residual.sugar \\
& + \beta_5 chlorides + \beta_6 free.sulfur.dioxide + \beta_7 total.sulfur.dioxide + \beta_8 density \\
& + \beta_9 pH + \beta_{10} sulphates + \beta_{11} alcohol + \varepsilon
\end{aligned}
$$

Fomula 2, Empirical Research into Both Wine
quality here resembles its sigmoid form

[APPENDIX III: Reference]

Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

Cortez P, António Cerdeira, Almeida F, et al. Modeling wine preferences by data mining from physicochemical properties[J]. Decision Support Systems, 2009, 47(4):547-553.

Biau, Gérard. Analysis of a Random Forests Model[J]. Journal of Machine Learning Research, 2010, 13(2):1063-1095.

Cutler K, Breiman L. Random forests[J]. Machine Learning, 2004, 45(1):157-176.

Schapire R E. A Brief Introduction to Boosting[C]. Sixteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc. 1999.

Zhu R L, Cao Y C, Pu Q M. Research on Information Technology with Wine Quality Evaluation Based on Neural Network[J]. Advanced Materials Research, 2014, 886:532-536.

James G, Witten D, Hastie T, et al. An Introduction to Statistical Learning[M]. Springer New York, 2013.

Keating K A, Cherry S. Use and Interpretation of Logistic Regression in Habitat-Selection[J]. Journal of Wildlife Management, 2011, 68(4):774-789.

Jaccard J, Turrisi R, Wan C K. Interaction effects in multiple regression[J]. Newbury Park California Sage Publications, 2003, 40(4):461.

[APPENDIX IV: Reproducible codes]