The Chemistry of Quality

White Wines of North Portugal

STATS 387 - Group 6: Alana Tauala, Leif Watkins, Sang Xing

Abstract

Despite the subjectivity of a concept like wine quality, the physicochemical measurements and quality scores of around 5,000 Vinho Verde white wines of Northern Portugal give insight into the best model to fit this type of data. Considering it as a classification problem, the team focused on prediction rather than inference and therefore not on the exact traits of a good wine. The decision boundary of the classification problem had a significant effect on the performance of the four models implemented. The QDA model for the dataset proves to be the superior of the four offering a high AUC value, low error rate, and balanced sensitivity and specificity values. Other methods were tested, suggesting a Bagging or Random Forest model to have the best performance.

❖ Introduction

Background

What gives a quality wine its quality? Objectively superior wine is fundamentally just a concept, one that has been in need of definition for many decades. Although quality is easy to detect, it is often difficult to define. Attempting to define an abstract idea such as quality relies heavily on subjective and extrinsic factors. Consequently, defining 'good' through chemistry can only ever be partially successful.

In an attempt to provide this definition, analyses were conducted on a dataset of Vinho Verde wines from Northern Portugal.[5] Vinho Verde translates literally to 'green wine' but is interpreted as 'young wine' as it is commonly released 3-6 months after the grapes are harvested. Often confused as a style of wine instead of a geographic location, Vinho Verde is made up of nine subregions in the northwest corner of Portugal, all the way up to the Spanish border. The landscape is one of rolling, green hillsides with two rivers, the Minho and Lima, flowing throughout. The geography of the country supports diverse microclimates, allowing for varying, unique wines. The vast majority of Vinho Verde wines are white and these will be the focus of all analyses.

Data, Cleaning & Mutation

Data Description

The dataset examined consists of 4898 observations of 11 physicochemical attributes and one output variable based on sensory data, of North Portugal's white "Vinho Verde" wine.[3] The physicochemical attributes were as follows:

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

The final variable, the response, was an integer with a range of 0-10 named 'quality'. The data was donated in October of 2009 and therefore may not reflect modern opinion.

Cleaning

To begin the cleaning process, the data set was checked for NA's or missing values in each observation. Upon finding no missing values, further exploration of the raw data revealed what appeared to be duplicate observations. The information source of the dataset states that due to privacy and logistic issues, there are no identifying characteristics included within the dataset. Therefore, the team could not confirm nor deny with any level of confidence if duplicate observations were caused by erroneous data input, accidental duplication, or if two different wines simply had equal physicochemical measurements. Due to this, the data was deduplicated to avoid contamination of the test data with training data, as well as to avoid supplying the model with artificial or inaccurate patterns as this can affect the performance of all models using this data for training and testing.

Mutation

The 'quality' response with a range 1-10 was mutated into a binary qualitative response. Observations with a quality value of 7 and above were labeled 'Good' and given a value of 1, all other observations were labeled 'Not Good' and given a value of 0. The 'good' variable was then converted into a factor for model creation and assessment.

Classification Methods

Classification

Classification algorithms predict discrete class labels, as opposed to a continuous quantity like regression methods. Although the dataset in its original form was fit for a regression method, the previously mentioned mutation of the response variable was intended to aid accuracy. It is believed that generalizing a continuous response as binary will provide more meaningful predictions in this context.[1]

❖ Methodology

K-nearest Neighbors (KNN) Classifier

The K-Nearest Neighbors Classification method (KNN) is a non-parametric, supervised learning classifier that uses proximity to make classifications or predictions on individual data points. Because it is non-parametric, meaning it makes no assumptions about distribution, one can expect it to outperform other methods such as LDA and Logistic Regression when the decision boundary is highly non-linear, given a large number of observations and a small number of predictors.[6] Most of the classifier algorithms are easy to implement for binary problems. However, only KNN has the ability to adjust to multi-class without any extra effort, making it much more versatile. [1]

Implementation

The first step to predicting the class of a new observation using the KNN classifier method is deciding the value of *K*. The choice of *K* can have a significant impact on the model's performance and may result in overfitting with too small of a *K* value, or underfitting with too large of a *K* value. The optimal *K* value will provide a minimized test error rate. The next step is to calculate the distance between the observation to be classified and all observations in the training set. Then, after selecting the *K* nearest neighbors based on the calculated distances, determine the majority class of the *K* nearest neighbors. Finally, the observation in question is then assigned to the majority class.

Starting with a *K*-value of 1, or *K*=1 neighbors, the team implemented 10-fold cross-validation. Using 9 of the 10 folds to build a model that is then used to make predictions on the last fold, this resampling method builds 10 separate models using a different combination of 9 folds as training data and records the error rate and other performance metrics to get averages. This method provides more accurate insight into how well a model may fit a dataset. This resampling method was implemented for each value of *K*, recording the error rate and other metrics at value from fold to calculate an average to refer to as the overall error rate for the associated value of *K* neighbors.

Logistic Regression (LR)

Logistic Regression is a supervised classification algorithm. Rather than modeling the response directly, LR models the probability that Y belongs to a particular category. The method works with both qualitative and quantitative response, for both inference and prediction, but is better suited for a binary qualitative response like the one here. Because it makes no distributional assumptions on predictors, it can perform better than other algorithms, like Linear Discriminant Analysis, when the Gaussian distributional assumptions are not met.[6]

The logistic regression method makes a few very critical assumptions about the data. The first assumption is that the relationship between the response and each predictor is linear. The second assumption is that there is no multicollinearity present in the linear combination of the predictors. A third assumption is of homoscedasticity, there should be constant variance in the predictors. The fourth assumption is of normality, the predictors should be normally distributed. If one of these assumptions is not met, the result of the logistic regression may be misleading or inaccurate. [1]

Implementation

Logistic regression algorithm can be expressed as:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

where, the left-hand side is called the logit, and $p(X)/\left(1-p(X)\right)$ is called the odds. The odds denote the ratio of probability of success to the probability of failure. So, in logistic regression, linear combinations of predictors are mapped to the $\log(odds)$ to predict the response of new observations. The logistic regression hypothesis is a generalization of the linear regression hypothesis. Using the sigmoid function, it creates an S-shaped decision boundary.

During the team's implementation of the logistic regression method, the data set was split into roughly 10 equal-sized folds. In this case, 70% of the total observations used for the training set were split here. For each fold, a logistic regression model was fit using the other 9 folds as training data. The fitted model was then used to predict the response for observations in the held-out fold. Accuracy, sensitivity, specificity, and AUC were then calculated for the predicted values of the held-out fold. Finally, averages of these values across the 10 folds were used as estimates of the model's performance on new data.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a supervised classification technique used for dimensionality reduction and data visualization. LDA works by projecting the data onto a lower-dimensional space, maximizing the separation between classes. It does this by calculating a set of linear discriminants that maximize the ratio of between-class variance to within-class variance. The method essentially finds the directions that best separate the different classes of data. It can perform better than some algorithms, especially logistic regression, when its assumptions are met.

LDA assumes that the data has a Gaussian distribution, which was not the case in this context, and that the covariance matrices of the different classes are equal. It also assumes that the data is linearly separable, meaning a linear decision boundary can accurately classify the different classes.

Implementation

During the LDA model fitting process, the method estimated the means and covariances of each input variable for the two classes (0 and 1). LDA used these estimates to construct a linear boundary between the two classes. Once the model was fitted, it was used to make predictions on new data by using Bayes' theorem to calculate the posterior probability of the input data belonging to each class and selecting the class with the highest probability. The decision boundary between the two classes is a linear combination function of the

input features. This can be visualized as a line in two-dimensional space or a plane in higher-dimensional space.

Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis is very much like its linear counterpart, with the exception that it allows for the different classes to have different variances.[4] This is one of the main differences between LDA and QDA and what makes QDA more flexible. QDA can be thought of as a compromise between the KNN method and the linear methods: LDA & logistic regression. Since the method assumes a quadratic decision boundary, it can accurately model a wider range of problems than the linear methods can. Though not as flexible as KNN, QDA can perform better with smaller training sets because of the decision boundary from assumptions.[6]

Implementation

Almost identical to the implementation of LDA, QDA uses estimates of means and covariance matrices of each class to construct a quadratic boundary between the two classes and calculate the likelihood and posterior probability of each class for new observations. The observation is classified as belonging to the class with the highest posterior probability exactly as in LDA. The major difference between the two methods is that QDA assumes that the covariance matrix for each class is different and estimates the mean and covariance matrix separately for each class.

❖ Results & Recommendations

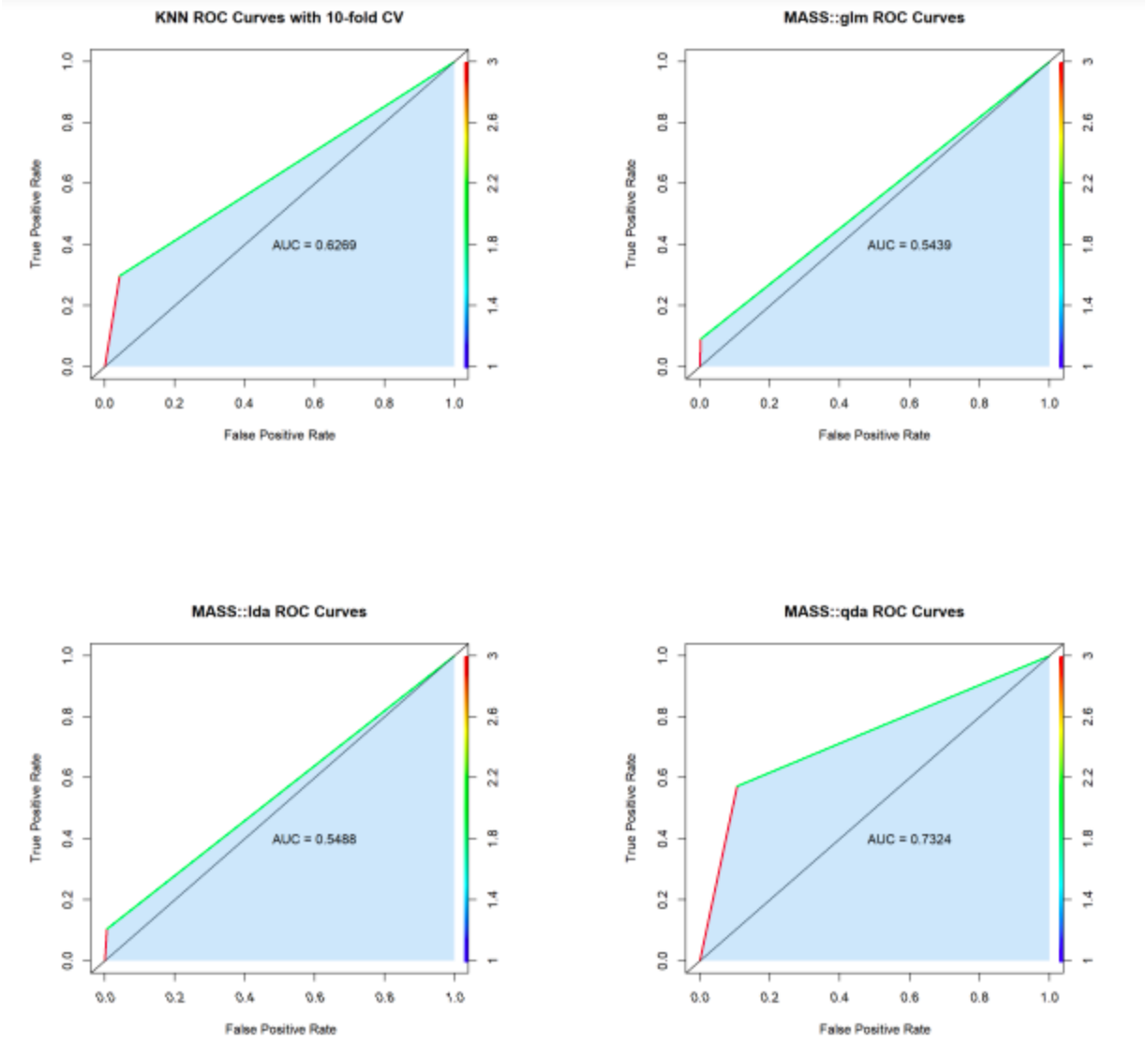
Model Assessments

The receiver operating characteristic (ROC) curve is a graph of the performance of a model at all classification thresholds. This curve plots two parameters: true positive rate and false positive rate.[1] Lowering the classification threshold causes the classification of more

items as positive, thus it would increase both false positives and true positives. The ‘perfect’ ROC curve is actually a 90-angle in the upper left corner.

The sensitivity of a model records the true positive rate. A high sensitivity indicates a model’s high ability to identify ‘good’ wines in this context. On the other hand, the specificity of a model records the true negative rate, or the model’s ability to correctly identify ‘not good’ wines.[2]

The AUC value stands for "Area under the Curve." This means AUC measures the entire two-dimensional area underneath the entire ROC curve. It is the most comprehensive assessment of model performance as it is an aggregate measure of performance across all possible classification thresholds.[1]



KNN

The KNN model had the highest error rate of the four models mentioned with a value of 0.1759. The high sensitivity of 0.9568, tells us that the model rarely misses a true good wine. A negative prediction from a model with such sensitivity holds weight and can be useful for 'ruling out' that that observation's response is positive. The model's specificity was 0.2971. A low specificity like this indicates that the model does not often correctly identify bad wines. Considering both measurements of sensitivity and specificity, the model seems to classify most wine it is presented with as good. With an AUC value of 0.62693, the model performed somewhat better than chance.

LR

The error rate of the logistic regression model was a favorable 0.1616. The sensitivity of the LR model was at its highest possible value of 1.000. A sensitivity of 100% indicates the model's ability to correctly predict every true positive or good wine when presented with it. Like the KNN model, a negative result can be very useful in many contexts. With a specificity of 0.0857, the lowest of all four models, the model has a very low ability to identify bad wines. Considering both of these measurements, it is safe to say the logistic regression model classifies almost every single wine it is presented with as good, but it also misclassifies a small proportion of bad wines as good. Its AUC value of 0.5438871, the lowest of all four methods mentioned, tells us that it hardly performs better than flipping a coin.

LDA

The LDA model had the lowest error rate of the models mentioned of 0.1591. With an extremely high sensitivity of 0.9969, the model consistently recognized a good wine. The slight specificity of 0.1143, is of no further benefit to the model's functionality. When considering both values, it is clear that, like the KNN and LR models, the LDA model classifies a very large majority of all observations as good. Unsurprisingly, the AUC value was 0.54881, indicating that LDA is not a particularly trustworthy model for this dataset. This is possibly due to the fact that the decision boundary in this case is not quite linear.

QDA

The QDA model had an error rate of 0.1692. With a satisfactory sensitivity level of 0.8934, one can reasonably trust the model to correctly identify a good wine. The sensitivity of the QDA model was the highest of all four models at 0.5714. This is also the smallest difference between sensitivity and specificity of all four models, indicating that it may be useful. It is possible that the method's assumptions have been met. The QDA model had the highest AUC value of 0.7324227, telling us that this model performed well across all thresholds.

Model	Resampling Method	Error Rate	Sensitivity	Specificity	AUC
KNN	10-fold CV (k=1:10)	0.1759	0.9568	0.2971	0.6269339
Logit	10-fold CV	0.1616	1.0000	0.0857	0.5438871
LDA	10-fold CV	0.1591	0.9969	0.1143	0.5488133
QDA	10-fold CV	0.1692	0.8934	0.5714	0.7324227

Results

During data analysis, model implementation, and performance assessment, the team was exposed to a few foreign concepts. The data insisted, according to the team's standard of a less than or equal to seven quality score being 'good' that nearly 80% of all Vinho Verde white wines are not good. The data also correctly expressed the negative correlation between residual sugars and alcohol, as well as how wine density and its insight into the quality of the fermentation process, have a linear effect on the quality score assigned to that observation.

The team recommends the QDA model for the dataset as the superior model for the following reasons: Although it had neither the lowest error rate, nor the highest sensitivity of the four models, the QDA model offered the highest AUC and specificity values out of all four models, strongly suggesting its superiority. To add to this suggestion, the QDA model also proved to have the most balanced specificity and sensitivity exhibiting the smallest difference across all other models' sensitivity/specificity measurements.

Conclusion

Although wine quality is a highly subjective concept, the chemistry and quality scores hold the power to give insight. Looking at this as a classification problem, the team was more focused on prediction than inference and therefore could not speak on the exact traits of a good wine. However, the team has a few recommendations for the best classification of newly released wine.

The QDA model is believed to perform the best because of the assumptions the method makes on the decision boundary of the classification problem in question. Other models were considered that insisted a Bagging model with all predictors and 3 number of trees or Random Forest model with all predictors and 500 number of trees would better predict new observations in this dataset with AUC values of 0.92196 and 0.91415 respectively.

❖ Works Cited

[1]: "Classification: ROC Curve and AUC." *Google Developers*, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:

[2]: Contributors to Wikimedia projects. "Sensitivity and Specificity." *Wikipedia*, 26 Jan. 2023, https://en.wikipedia.org/wiki/Sensitivity_and_specificity.

[3]: Cortez, Paulo, et al. "Modeling Wine Preferences by Data Mining from Physicochemical Properties." *Decision Support Systems*, vol. 47, no. 4, Nov. 2009, pp. 547–53, https://doi.org/10.1016/j.dss.2009.05.016.

[4]: Sohil, Fariha, et al. "An Introduction to Statistical Learning with Applications in R." *Statistical Theory and Related Fields*, vol. 6, no. 1, Sept. 2021, pp. 87–87, https://doi.org/10.1080/24754269.2021.1980261.

[5]: "Wine Quality - an Overview." *ScienceDirect Topics*, https://www.sciencedirect.com/topics/food-science/wine-quality.

[6]: Yan, Nancy. "Study Note: Comparing Logistic Regression, LDA, QDA, and KNN." *Nancy's Notes*, https://nancyyanyu.github.io/posts/6084c2b2/. Accessed 20 Mar. 2023.