Classifying Wine Quality

Kabir Kedia, Manjot Nagyal, Elizabeth Ouanemalay, Kenedy Sanchez, Nilsu Uzunlar

Carnegie Mellon University

Introduction

The focus of this project is to identify if physicochemical properties of wine can reliably predict perceived wine quality. We applied five classification models,Logistic Regression, Pruned Decision Trees, Random Forest, XGBoost and KNN , to study the binary division of wine quality as `GOOD` or `BAD`. Using 6,497 data points, 11 predictor variables and 1 response variable the model with the highest Area Under Curve (AUC) was selected for determining which properties impacted perceived wine quality the most.

Exploratory Data Analysis

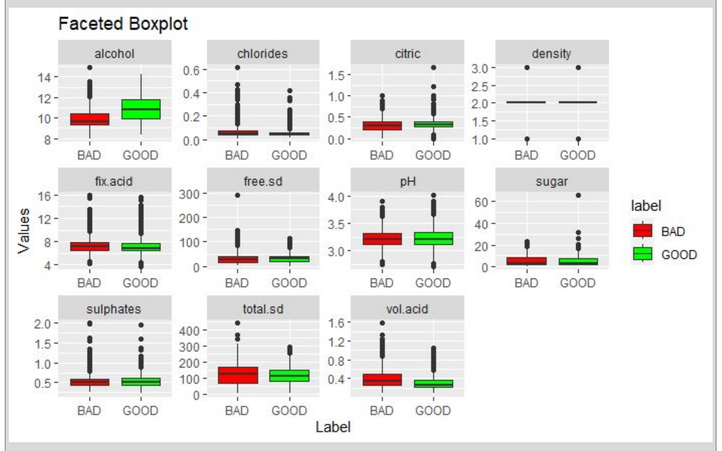

Figure 1 : Faceted boxplots of physicochemical property distributions against response variable

The sugar values (greater than 20), citric (greater than 1.0) chlorides (greater than 0.2), free sulfate (greater than 275) contained outliers that were removed prior to applying the classification methods. 431 rows were removed

Classification Methods and Results

Model Comparison

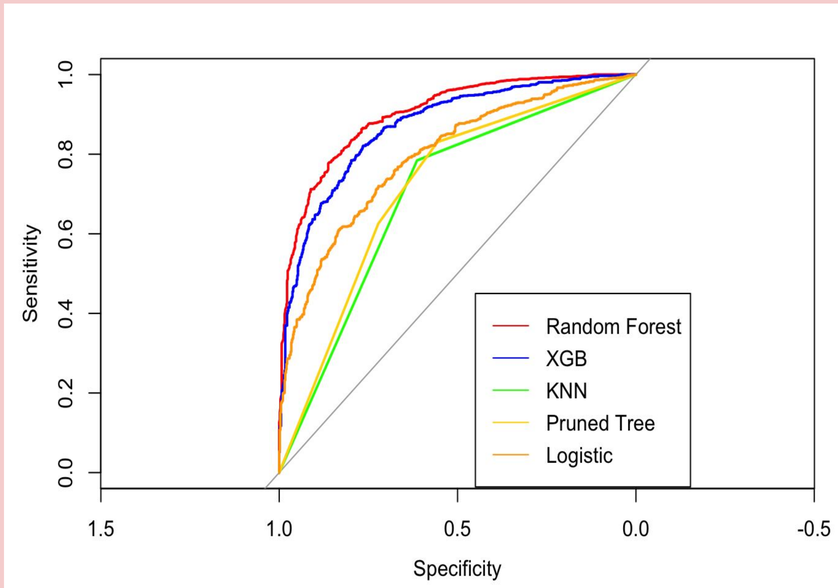

Figure 2 : ROC curves for different classification models

MODEL	AUC	MCR
XGBoost	0.87	0.198
Logistic Regression	0.794	0.267
Logistic Regression BSS	0.792	0.31
KNN	0.699	0.278
Random Forest	0.9	0.192
Pruned Tree	0.717	0.269

Table 1: AUC and MCR values

Among the five classification models, random forest computed the lowest MCR 19.2% and the highest AUC value suggesting that Random Forest is the best model for classifying the wine quality.

When AUC values are compared for Logistic regression and BSS models (under BIC) , it is observed that AUC for Logistic regression (0.794) is greater than that of BSS (0.792). Hence, proceeding on with the best subset hurts our classification performance; albeit slightly.

Best Model - Random Forest

Prior to calculating these model statistics, the Youden's J Index of 0.52 was applied to segregate between the binary classes. Models took into account properties such as `pH, acidity, and alcohol content,` showing wine quality is driven by multiple factors. Though all classification models used the same predictor variables random forest classified the `GOOD` and `BAD` labeling most accurately.

	BAD	GOOD
BAD	509	231
GOOD	81	803

Table 2: Confusion matrix for Random Forest

Comparison of Results: Logistic Regression Best Subset Selection (BSS) vs. Variable Importance Plot (VIP)

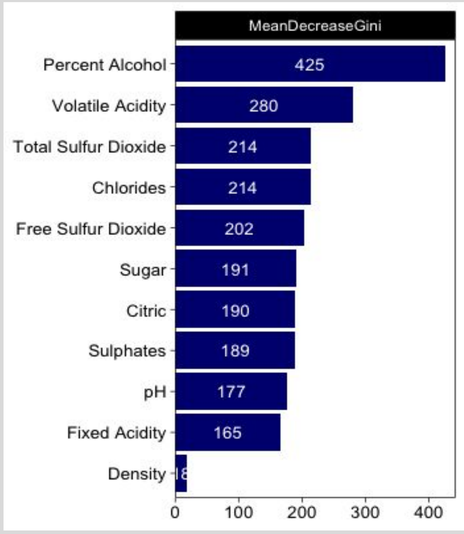

Figure 3: Variable Importance for Random Forest

VIP shows that `percent alcohol` is the best predictor variable for classification. According to the plot, `volatile acidity, total sulfur dioxide, chlorides, free sulfur oxide, and sugar` are the other important variables. Logistic regression best subset selection (BSS) reveals that `volatile acidity, sugar, free sulfur dioxide, total sulfur dioxide, sulphates, and percent alcohol` are important for classifying the wine quality. When both models are compared, VIP selects chlorides as an important feature; whereas BSS selects sulphates as one of the variables that stays in the best subset instead of chlorides.