STA467 Final Project

Allen Liu

2024-04-03

I. Introduction

This wine data set presents an intriguing opportunity for predictive modeling, focusing on physicochemical properties and sensory data of the red wine variants of the Portuguese “Vinho Verde” wine. Sourced from the UCI machine learning repository and detailed by [Cortez et al., 2009], this data set offers a classification task: determining wine quality based on a 0 to 10 scale.

Logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), regularization techniques (Lasso, Ridge, Elastic Net), and random forest will be used to analyze this data set. Each model is tuned and evaluated using repeated cross-validation techniques for robustness and accuracy.

By setting a binary classification threshold for wine quality and leveraging advanced modeling approaches, the goal is to uncover the physiochemical attributes that differentiate red wines. The focus of this analysis extends beyond prediction; the aim is to understand the interaction of variables contributing to wine quality perception, thereby contributing to enology and predictive modeling.

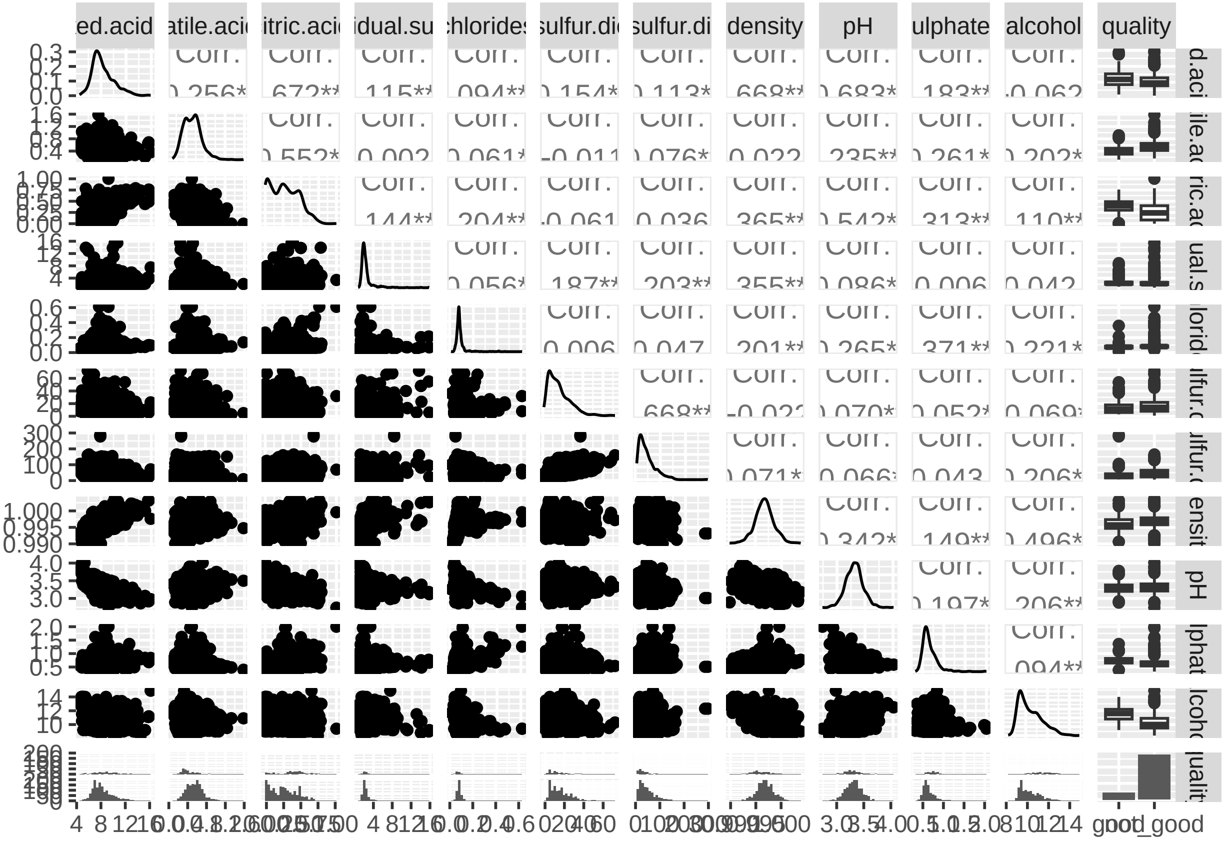
II. Exploratory Data Analysis (EDA)

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##    quality
## 1 not_good
## 2 not_good
## 3 not_good
## 4 not_good
## 5 not_good
## 6 not_good
```

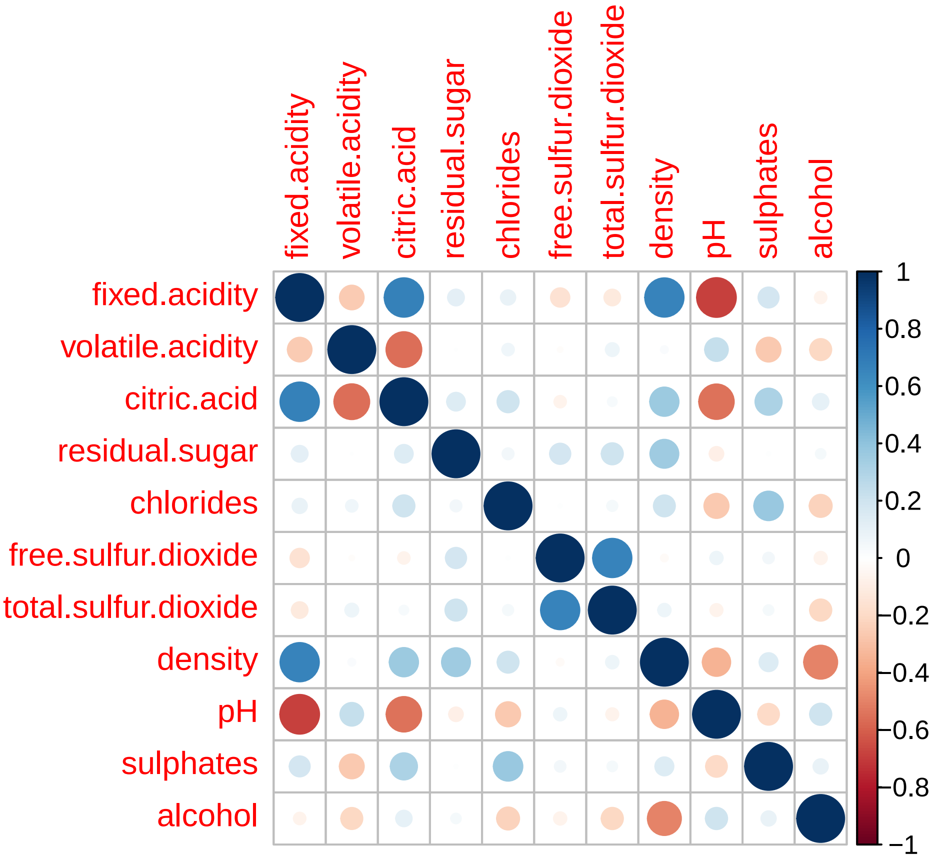
As part of the exploratory data analysis (EDA), the ‘quality’ variable will be modified into a binary format. This modification involves categorizing wines into two distinct groups: ‘good’ and ‘not good.’ The rationale behind this transformation is to simplify the analysis and modeling tasks by focusing on whether a wine is considered ‘good’ rather than its specific quality score.

An arbitrary cutoff point will be set at a quality score of 7 or higher, classifying wines with scores above this threshold as ‘good’ and the remaining wines as ‘not good.’ This decision is informed by domain knowledge and prior research indicating that wines with higher quality scores are generally perceived more favorably by consumers.

By converting the ‘quality’ variable into a binary format, it facilitates the identification of key factors that contribute to the perception of wine quality. Throughout this analysis, the binary ‘quality’ variable will be referred to, investigating the factors that distinguish ‘good’ wines from ‘not good’ ones, providing valuable insights for enology and predictive modeling tasks.



From the ggpairs plot, some of the predictors exhibit right-skewed or non-normally distributed patterns. This observation is particularly notable in variables such as ‘residual sugar,’ ‘chlorides,’ ‘free sulfur dioxide,’ ‘total sulfur dioxide,’ and ‘sulphates.’ The right-skewed nature of these variables indicates a higher frequency of lower values with a tail stretching towards higher values. This skewness can impact the performance of certain statistical models that assume a normal distribution of data, potentially leading to biased estimates or inaccurate predictions. Therefore, the skewness in these predictors will be addressed through modeling techniques that can handle non-normal data effectively. This ensures that the modeling process accounts for the distributional characteristics of the predictors, ultimately enhancing the accuracy and reliability of the predictive models.



Fixed Acidity vs. Citric Acid: There is a strong positive correlation (approximately 0.67) between fixed acidity and citric acid. This indicates that wines with higher fixed acidity tend to have higher levels of citric acid as well.

Fixed Acidity vs. Density: Fixed acidity also shows a moderately positive correlation (around 0.67) with density. Wines with higher fixed acidity may thus tend to have higher densities.

Volatile Acidity vs. Citric Acid: There is a moderate negative correlation (about -0.55) between volatile acidity and citric acid. Wines with higher levels of volatile acidity are likely to have lower levels of citric acid.

pH vs. Fixed Acidity and Citric Acid: pH exhibits a strong negative correlation with fixed acidity (around -0.68) and a moderate negative correlation with citric acid (about -0.54). This suggests that wines with higher fixed acidity and lower citric acid content tend to have lower pH levels.

Alcohol vs. Density: Alcohol content shows a moderate negative correlation (approximately -0.50) with density. Wines with higher alcohol content may have lower densities.

It's important to note that while certain variables may exhibit strong correlations, the models used for analysis will still incorporate the full predictor set initially. This approach ensures that the models consider all available information and relationships among the predictors before any feature selection or removal is performed. Later in the analysis, the impact of removing predictors based on correlations or other criteria will be explored, and the performance of models with and without certain predictors will be compared (**refer to Appendix B**). This comparative analysis will provide insights into the importance of individual predictors and their contribution to predictive modeling.

III. Modeling Approach, Building, and Evaluation

The modeling approach employed for predicting wine quality involved utilizing a variety of machine learning algorithms. Specifically, the following models were trained and evaluated:

1. Logistic Regression:

- The logistic regression model was trained using the "glm" method with repeated cross-validation (CV) performed using 10 folds and 10 repeats. The data was preprocessed by centering and scaling.

2. Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA):

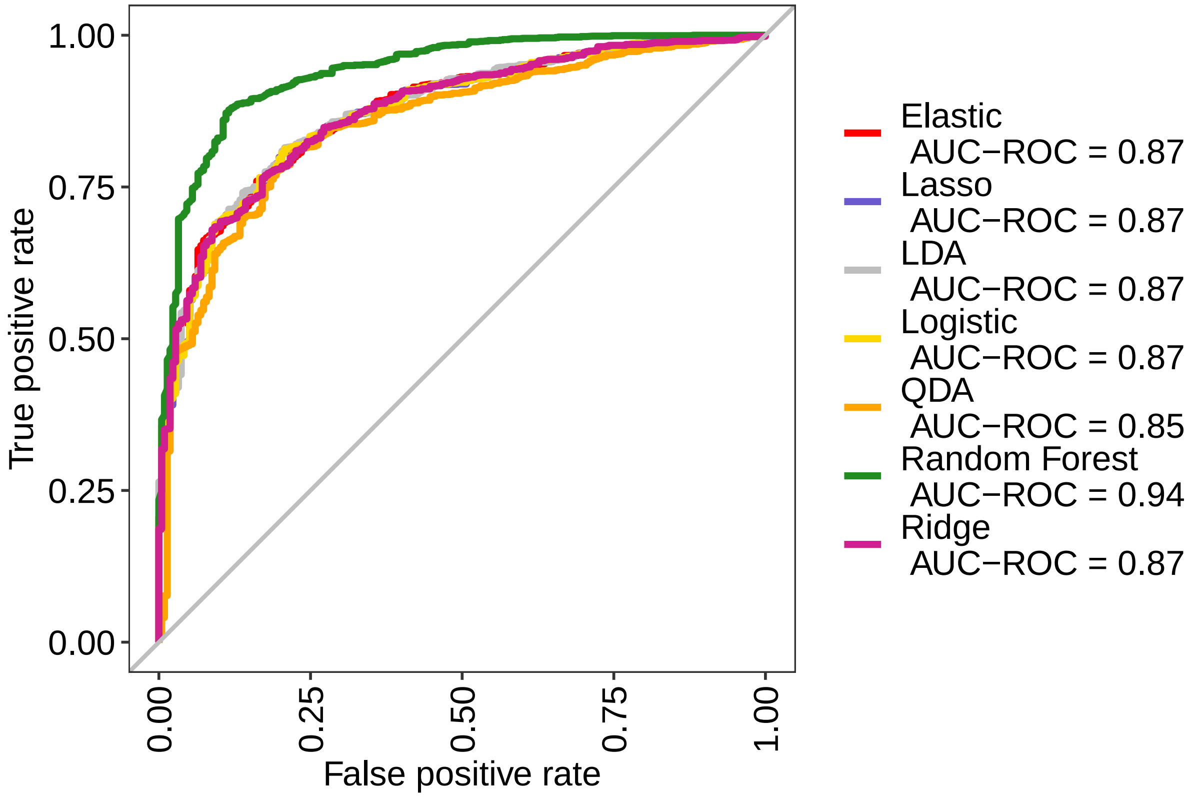
- LDA and QDA models were trained using their respective methods with the same repeated CV setup and preprocessing as logistic regression.

3. Lasso, Ridge, and Elastic Net Regression:

- Lasso, Ridge, and Elastic Net models were trained using the "glmnet" method with repeated CV and preprocessing similar to the other models. Different regularization parameters (alpha and lambda) were tuned to optimize model performance.

4. Random Forest:

- The Random Forest model was trained using the "rf" method with repeated CV, tuning the number of variables randomly sampled as candidates at each split (mtry) and setting the number of trees (ntree) to 100.



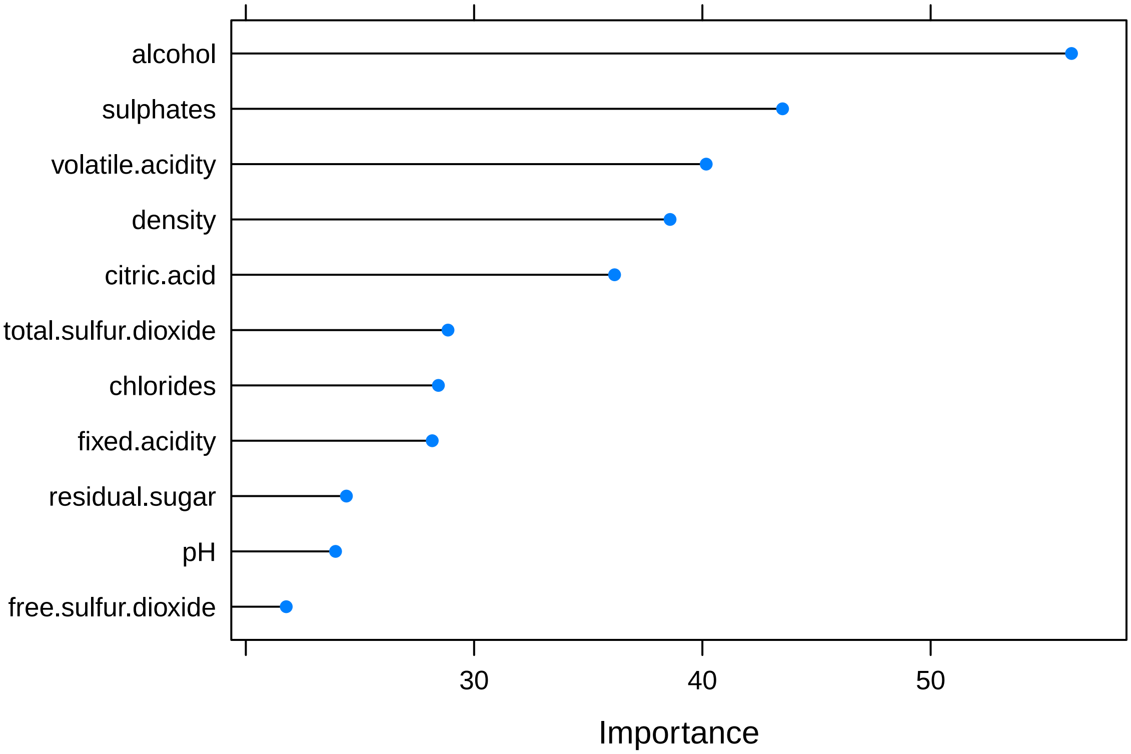
```
##    Resample          Elastic~ROC      Elastic~Sens      Elastic~Spec
##  Length:100         Min.   :0.7839   Min.   :0.04762   Min.   :0.9420
##  Class :character   1st Qu.:0.8552   1st Qu.:0.17208   1st Qu.:0.9783
##  Mode  :character   Median :0.8723   Median :0.22727   Median :0.9855
##                     Mean   :0.8747   Mean   :0.23377   Mean   :0.9821
##                     3rd Qu.:0.9024   3rd Qu.:0.28571   3rd Qu.:0.9928
##                     Max.   :0.9538   Max.   :0.47619   Max.   :1.0000
##    Lasso~ROC        Lasso~Sens        Lasso~Spec        LDA~ROC
##  Min.   :0.7795   Min.   :0.04545   Min.   :0.9058   Min.   :0.7847
##  1st Qu.:0.8502   1st Qu.:0.27273   1st Qu.:0.9565   1st Qu.:0.8543
##  Median :0.8775   Median :0.31818   Median :0.9638   Median :0.8729
##  Mean   :0.8727   Mean   :0.32338   Mean   :0.9644   Mean   :0.8750
##  3rd Qu.:0.8947   3rd Qu.:0.36797   3rd Qu.:0.9783   3rd Qu.:0.8993
##  Max.   :0.9496   Max.   :0.54545   Max.   :0.9928   Max.   :0.9500
##     LDA~Sens          LDA~Spec      Logistic~ROC    Logistic~Sens
##  Min.   :0.1364   Min.   :0.8986   Min.   :0.7795   Min.   :0.04545
##  1st Qu.:0.3182   1st Qu.:0.9420   1st Qu.:0.8497   1st Qu.:0.27273
##  Median :0.3723   Median :0.9565   Median :0.8775   Median :0.31818
##  Mean   :0.3805   Mean   :0.9528   Mean   :0.8726   Mean   :0.32478
##  3rd Qu.:0.4545   3rd Qu.:0.9658   3rd Qu.:0.8951   3rd Qu.:0.38095
##  Max.   :0.7273   Max.   :1.0000   Max.   :0.9500   Max.   :0.57143
##  Logistic~Spec       QDA~ROC          QDA~Sens          QDA~Spec
##  Min.   :0.8986   Min.   :0.7057   Min.   :0.3182   Min.   :0.7971
##  1st Qu.:0.9565   1st Qu.:0.8244   1st Qu.:0.4708   1st Qu.:0.8696
##  Median :0.9638   Median :0.8516   Median :0.5455   Median :0.8986
##  Mean   :0.9644   Mean   :0.8519   Mean   :0.5513   Mean   :0.8933
##  3rd Qu.:0.9783   3rd Qu.:0.8851   3rd Qu.:0.6364   3rd Qu.:0.9130
##  Max.   :0.9928   Max.   :0.9558   Max.   :0.8182   Max.   :0.9783
##  Random Forest~ROC Random Forest~Sens Random Forest~Spec   Ridge~ROC
##  Min.   :0.8699    Min.   :0.2857     Min.   :0.9275     Min.   :0.7816
##  1st Qu.:0.9215    1st Qu.:0.4545     1st Qu.:0.9710     1st Qu.:0.8482
##  Median :0.9385    Median :0.5000     Median :0.9783     Median :0.8757
##  Mean   :0.9337    Mean   :0.5243     Mean   :0.9777     Mean   :0.8738
##  3rd Qu.:0.9513    3rd Qu.:0.5763     3rd Qu.:0.9855     3rd Qu.:0.9017
##  Max.   :0.9725    Max.   :0.7727     Max.   :1.0000     Max.   :0.9548
##    Ridge~Sens         Ridge~Spec
##  Min.   :0.04762   Min.   :0.9420
##  1st Qu.:0.21807   1st Qu.:0.9640
##  Median :0.27273   Median :0.9783
##  Mean   :0.27634   Mean   :0.9758
##  3rd Qu.:0.33333   3rd Qu.:0.9855
##  Max.   :0.54545   Max.   :1.0000
```

IV. Results, Discussion, and Conclusion

Results

1. Model Performance

The ROC curves reveal that the Random Forest model outperforms all others, demonstrating its superior fit for the wine data set. On the other hand, the Quadratic Discriminant Analysis (QDA) exhibits the lowest ROC curve, indicating comparatively weaker predictive performance. Among the remaining models, including Elastic Net, Lasso, LDA, Logistic Regression, and Ridge Regression, their ROC curves are closely clustered with negligible differences in performance. This suggests that while these models are competitive, the Random Forest model stands out as the most effective choice for accurately discriminating between wine quality.



2. Feature Importance Analysis

The variable importance analysis reveals key insights into the importance of physicochemical features in predicting wine quality. Among the top-ranking features, alcohol is the most influential predictor with a high importance score of 57.95. This is followed closely by sulphates (44.12), volatile acidity (38.70), density (36.52), and citric acid (34.98). These features significantly contribute to the Random Forest model's ability to accurately classify wines as good or not good.

Discussion

1. Interpretation of Feature Importance

The high importance of alcohol, sulphates, volatile acidity, density, and citric acid highlights their crucial role in determining wine quality. These features likely capture aspects related to flavor profile, acidity levels, and alcohol content, which are known factors influencing wine quality (Understanding Acidity in Wine, 2024).

2. Specificity and Sensitivity Specificity and sensitivity are vital metrics in classification tasks. Specificity measures the ability of a model to correctly identify negative instances, indicating how well it avoids false positives. On the other hand, sensitivity quantifies the model's ability to detect positive instances accurately without missing actual positive cases, i.e. how well it avoids false negatives.

3. Across the models tested: Random Forest exhibited a high sensitivity of up to 0.77, indicating its effectiveness in correctly identifying positive wine quality instances. Specificity values were generally high across models, with Random Forest showing notable performance in correctly identifying negative instances. Sensitivity values, on the other hand, were generally low across models.

Conclusion

1. Key Findings

The Random Forest model, driven by key features such as alcohol, sulphates, volatile acidity, density, and citric acid, emerged as the top-performing model for wine quality prediction. Understanding the importance of these features provides important insights for wine producers and industry stakeholders.

In analyzing the low sensitivity values observed across some models, it becomes apparent that these models are better at predicting wines of lower quality (mediocre to low quality) compared to identifying wines of high quality. This phenomenon suggests that the models may excel at detecting negative instances (e.g., poor-quality wines) but struggle to identify positive instances (e.g., good-quality wines) with the same level of accuracy. One potential factor contributing to this imbalance in predictive performance is the nature of the dataset itself, which may be skewed towards containing more instances of mediocre to low-quality wines than instances of high-quality wines. This imbalance can lead to a higher emphasis on learning patterns associated with negative instances, thereby affecting the models' ability to generalize well to positive instances.

2. Practical Implications

The variable importance analysis has practical implications for wine production and quality improvement strategies. Producers can use these insights to optimize wine formulations, enhance quality control measures, and tailor products to meet consumer preferences effectively.

In the context of wine quality assessment, where the focus often lies on identifying exceptional or high-quality wines, the low sensitivity values raise concerns about the models' effectiveness in precisely classifying such instances. Therefore, while the models may exhibit strong performance in terms of specificity (identifying non-good wines correctly), their lower sensitivity indicates a potential limitation in capturing and accurately predicting instances of high wine quality.

3. Future Directions

Future research endeavors may focus on exploring additional predictors or refining modeling techniques to further enhance predictive accuracy and deepen understanding of wine quality determinants. Future research into specific qualities of wines could look into tannin content, alcohol content, or acidity to narrow the scope (Dufourc, 2021).

Furthermore, addressing the challenge of having low sensitivities across the models requires strategies such as:

1. **Balancing the Dataset:** Collecting additional data or employing sampling techniques to balance the representation of different wine quality categories can help mitigate the effects of class imbalance.
2. **Adjusting Model Parameters:** Fine-tuning model parameters, such as adjusting class weights or using algorithms specifically designed for imbalanced datasets, can improve the models' sensitivity towards positive instances.
3. **Feature Engineering:** Incorporating domain knowledge and relevant features that better capture the characteristics of high-quality wines can enhance the models' ability to identify and predict such instances accurately.

By acknowledging and addressing these considerations, future iterations of the analysis can aim to improve the models' sensitivity specifically towards wines of high quality, aligning more closely with the practical objectives of wine quality assessment and decision-making in the industry.

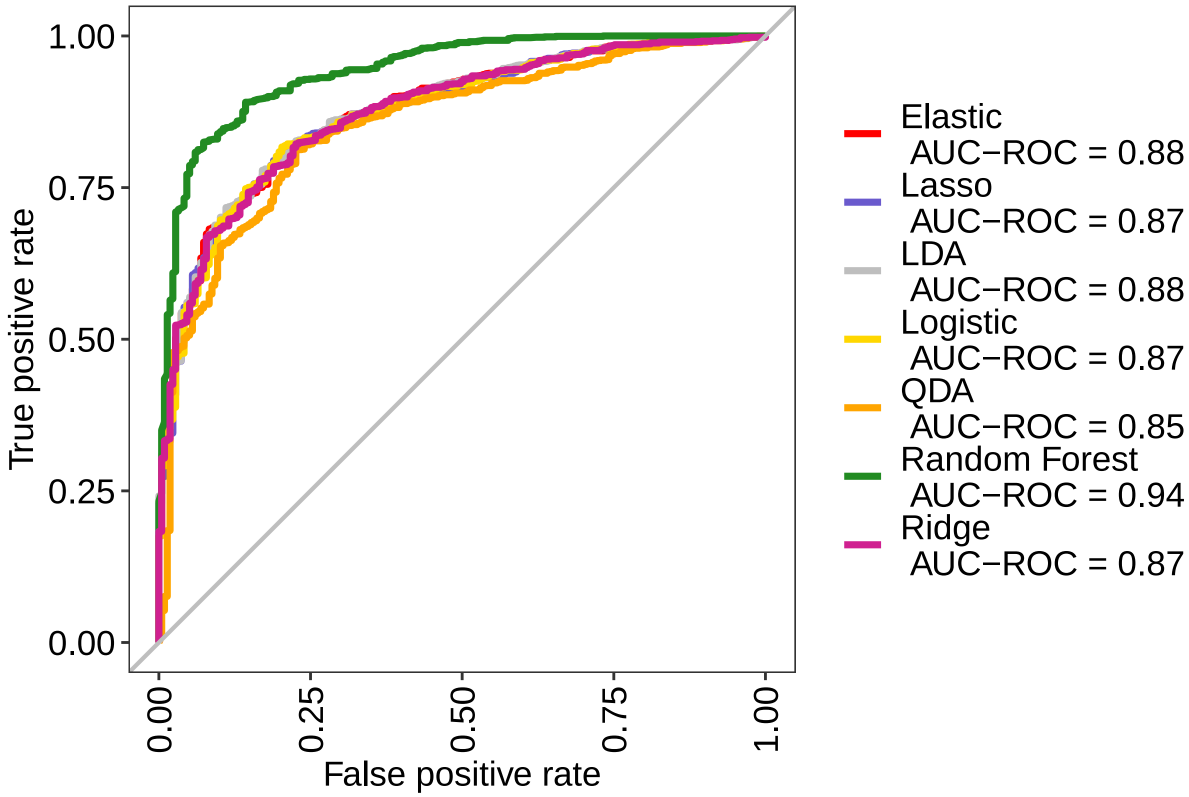
V. References and Appendices

References

Dufourc, J. E., (2021). Wine tannins, saliva proteins and membrane lipids. Biochimica et Biophysica Acta (BBA)- Biomembranes. 1863(10). Understanding Acidicty in Wine. (2024). Understanding acidity in wine. Wine Folly. 2024.https://winefolly.com/deep-dive/understanding-acidity-in-wine/

UCI Machine Learning. “Red Wine Quality.” Kaggle, UCI, 27 Nov. 2017, www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/data.

Appendix A: Sensitivity Improvement After Upsampling



```
##    Resample          Elastic~ROC      Elastic~Sens     Elastic~Spec
##  Length:100         Min.   :0.7581   Min.   :0.5714   Min.   :0.6667
##  Class :character   1st Qu.:0.8538   1st Qu.:0.7619   1st Qu.:0.7536
##  Mode  :character   Median :0.8789   Median :0.8182   Median :0.7826
##                     Mean   :0.8748   Mean   :0.8092   Mean   :0.7794
##                     3rd Qu.:0.9017   3rd Qu.:0.8636   3rd Qu.:0.8043
##                     Max.   :0.9522   Max.   :0.9545   Max.   :0.8561
##    Lasso~ROC         Lasso~Sens       Lasso~Spec         LDA~ROC
##  Min.   :0.7708   Min.   :0.5909   Min.   :0.7029   Min.   :0.7941
##  1st Qu.:0.8524   1st Qu.:0.7619   1st Qu.:0.7609   1st Qu.:0.8587
##  Median :0.8760   Median :0.8182   Median :0.7826   Median :0.8803
##  Mean   :0.8732   Mean   :0.8126   Mean   :0.7842   Mean   :0.8763
##  3rd Qu.:0.8991   3rd Qu.:0.8636   3rd Qu.:0.8072   3rd Qu.:0.8986
##  Max.   :0.9693   Max.   :1.0000   Max.   :0.8768   Max.   :0.9335
##     LDA~Sens         LDA~Spec       Logistic~ROC    Logistic~Sens
##  Min.   :0.6190   Min.   :0.6978   Min.   :0.7862   Min.   :0.5714
##  1st Qu.:0.7727   1st Qu.:0.7536   1st Qu.:0.8542   1st Qu.:0.7727
##  Median :0.8182   Median :0.7754   Median :0.8743   Median :0.8182
##  Mean   :0.8257   Mean   :0.7735   Mean   :0.8722   Mean   :0.8147
##  3rd Qu.:0.8636   3rd Qu.:0.7971   3rd Qu.:0.8940   3rd Qu.:0.8636
##  Max.   :1.0000   Max.   :0.8696   Max.   :0.9441   Max.   :0.9545
##  Logistic~Spec       QDA~ROC          QDA~Sens         QDA~Spec
##  Min.   :0.6835   Min.   :0.7484   Min.   :0.6190   Min.   :0.6594
##  1st Qu.:0.7609   1st Qu.:0.8158   1st Qu.:0.7619   1st Qu.:0.7174
##  Median :0.7842   Median :0.8470   Median :0.8182   Median :0.7464
##  Mean   :0.7832   Mean   :0.8481   Mean   :0.8125   Mean   :0.7433
##  3rd Qu.:0.8116   3rd Qu.:0.8858   3rd Qu.:0.8636   3rd Qu.:0.7681
##  Max.   :0.8841   Max.   :0.9396   Max.   :1.0000   Max.   :0.8333
##  Random Forest~ROC Random Forest~Sens Random Forest~Spec   Ridge~ROC
##  Min.   :0.8539    Min.   :0.3333     Min.   :0.9203     Min.   :0.7740
##  1st Qu.:0.9169    1st Qu.:0.5000     1st Qu.:0.9496     1st Qu.:0.8524
##  Median :0.9408    Median :0.5909     Median :0.9638     Median :0.8775
##  Mean   :0.9359    Mean   :0.5828     Mean   :0.9631     Mean   :0.8746
##  3rd Qu.:0.9574    3rd Qu.:0.6667     3rd Qu.:0.9711     3rd Qu.:0.8978
##  Max.   :0.9874    Max.   :0.8095     Max.   :0.9928     Max.   :0.9657
##    Ridge~Sens       Ridge~Spec
##  Min.   :0.5000   Min.   :0.6812
##  1st Qu.:0.7619   1st Qu.:0.7523
##  Median :0.8182   Median :0.7790
##  Mean   :0.8070   Mean   :0.7794
##  3rd Qu.:0.8636   3rd Qu.:0.7986
##  Max.   :0.9545   Max.   :0.8921
```

Before Upsampling

Before upsampling, the ROC curve values for most models indicated moderate to high performance, with minimal variation observed. Models like Ridge and Elastic showed slight improvements in their ROC curves after upsampling, indicating enhanced overall predictive power. However, the ROC curves of other models remained relatively stable, suggesting that their discriminatory ability between positive and negative instances did not significantly change.

After Upsampling

Despite the limited improvement in ROC curve values post-upsampling, there was a substantial enhancement in sensitivity across various models. This improvement is particularly crucial as it signifies a significant boost in the models' ability to correctly identify positive instances, such as high-quality wines. The upsampling technique effectively addressed the imbalance in the dataset, allowing the models to better capture the minority class and make more accurate predictions for positive cases.

Random Forest Exception

The Random Forest model, known for its ability to handle class imbalances effectively, showed consistent ROC curve values before and after upsampling. While the ROC curve did not demonstrate significant improvement, the model's sensitivity, although not dramatically enhanced, still benefited from the upsampling process. This indicates that Random Forest maintained its overall predictive power even with the dataset adjustments.

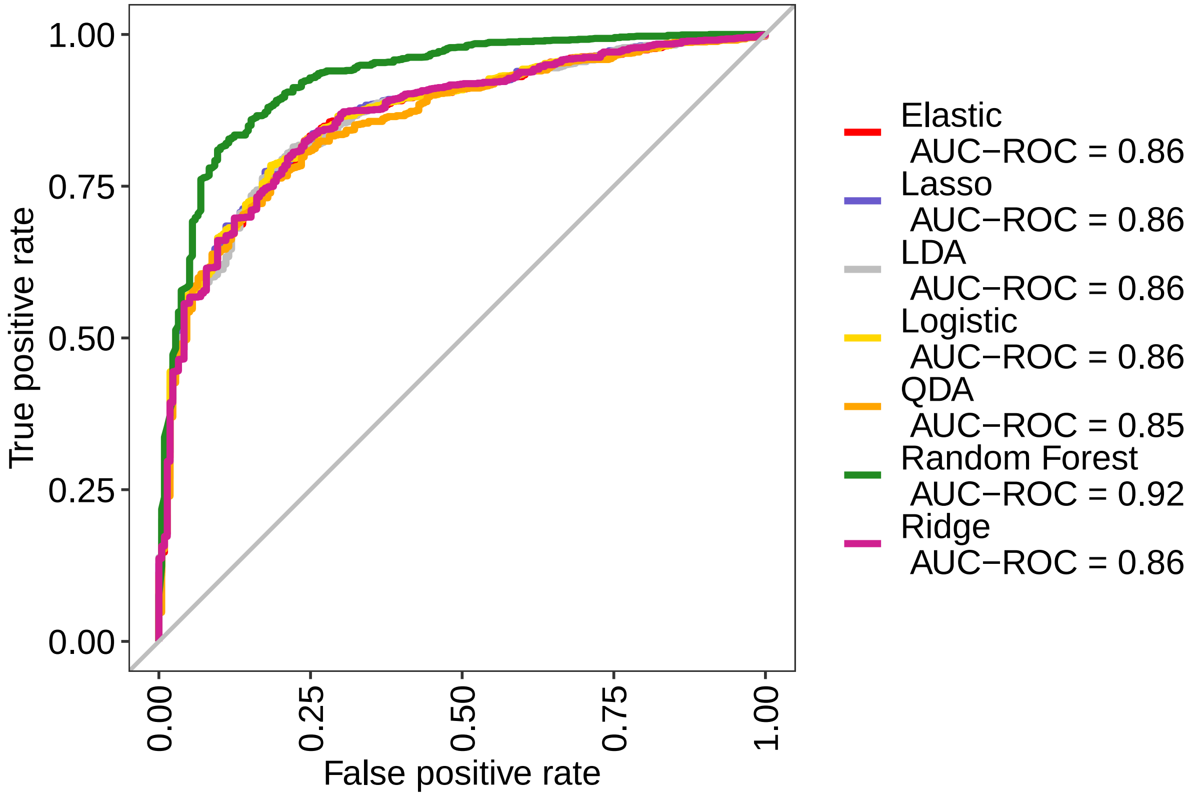
Importance of Sensitivity Improvement

The notable improvement in sensitivity post-upsampling holds more significant practical implications than the modest changes in ROC curve values. Sensitivity directly influences the models' ability to detect positive instances accurately, aligning with the primary objective of identifying high-quality wines. Therefore, the sensitivity improvement observed after upsampling reinforces the effectiveness of this technique in improving model performance for imbalanced datasets.

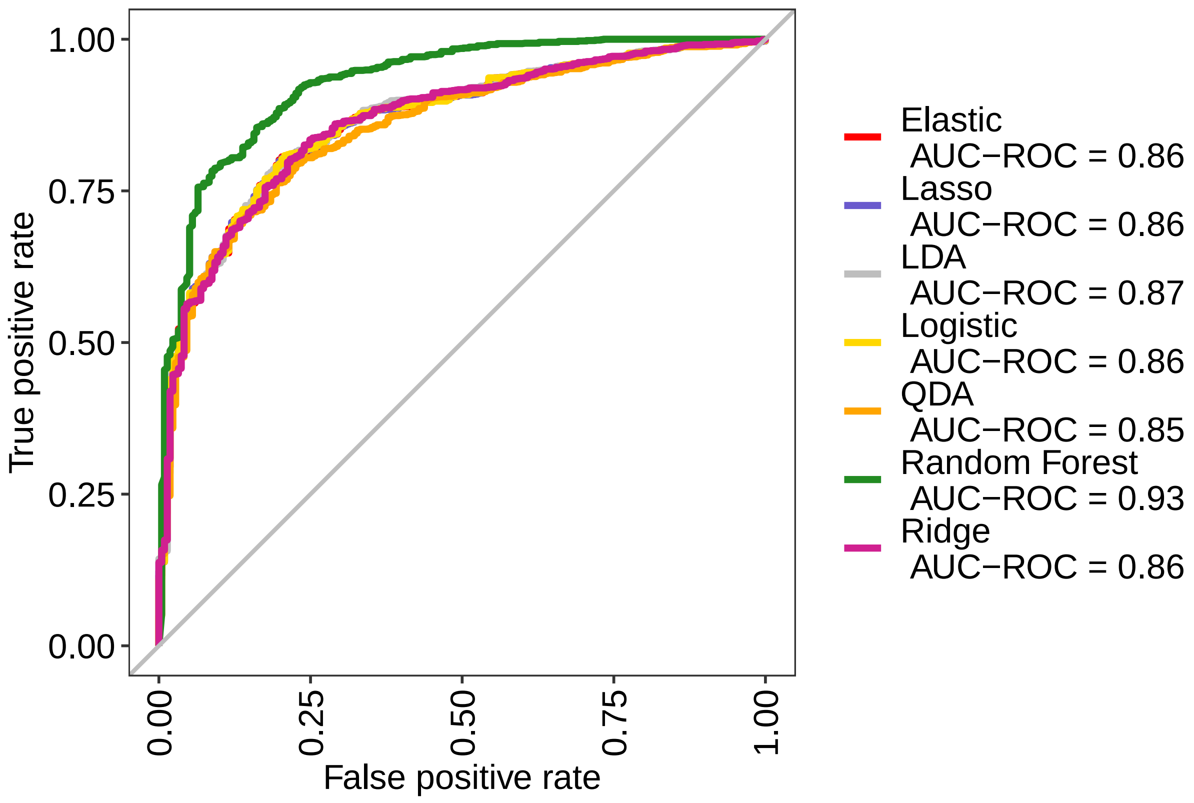
Considerations

While sensitivity improvement is a positive outcome of upsampling, it's essential to consider potential trade-offs and drawbacks. Upsampling can introduce biases or noise into the data, leading to overfitting or reduced generalization performance. Additionally, upsampling increases computational complexity and resource requirements, which may impact model scalability and real-time prediction capabilities. Hence, a balanced approach incorporating various validation techniques and performance metrics is crucial to evaluate model effectiveness comprehensively.

Appendix B: Subset of Predictors Analysis



```
##    Resample          Elastic~ROC       Elastic~Sens       Elastic~Spec
##  Length:100          Min.   :0.7750    Min.   :0.04545    Min.   :0.9275
##  Class :character    1st Qu.:0.8323    1st Qu.:0.19048    1st Qu.:0.9565
##  Mode  :character    Median :0.8714    Median :0.27273    Median :0.9710
##                      Mean   :0.8643    Mean   :0.25996    Mean   :0.9661
##                      3rd Qu.:0.8931    3rd Qu.:0.31818    3rd Qu.:0.9783
##                      Max.   :0.9337    Max.   :0.52381    Max.   :1.0000
##    Lasso~ROC        Lasso~Sens        Lasso~Spec        LDA~ROC
##  Min.   :0.7931   Min.   :0.04545   Min.   :0.9203   Min.   :0.7810
##  1st Qu.:0.8343   1st Qu.:0.14286   1st Qu.:0.9710   1st Qu.:0.8351
##  Median :0.8713   Median :0.19048   Median :0.9783   Median :0.8635
##  Mean   :0.8654   Mean   :0.21056   Mean   :0.9789   Mean   :0.8624
##  3rd Qu.:0.8919   3rd Qu.:0.27273   3rd Qu.:0.9856   3rd Qu.:0.8899
##  Max.   :0.9413   Max.   :0.40909   Max.   :1.0000   Max.   :0.9372
##     LDA~Sens         LDA~Spec       Logistic~ROC    Logistic~Sens
##  Min.   :0.1364   Min.   :0.8986   Min.   :0.7450   Min.   :0.0000
##  1st Qu.:0.2727   1st Qu.:0.9353   1st Qu.:0.8411   1st Qu.:0.2273
##  Median :0.3333   Median :0.9531   Median :0.8613   Median :0.2381
##  Mean   :0.3226   Mean   :0.9520   Mean   :0.8647   Mean   :0.2603
##  3rd Qu.:0.3810   3rd Qu.:0.9710   3rd Qu.:0.8917   3rd Qu.:0.3182
##  Max.   :0.5455   Max.   :0.9928   Max.   :0.9386   Max.   :0.5455
##  Logistic~Spec       QDA~ROC          QDA~Sens         QDA~Spec
##  Min.   :0.9275   Min.   :0.7740   Min.   :0.1364   Min.   :0.8841
##  1st Qu.:0.9565   1st Qu.:0.8257   1st Qu.:0.3182   1st Qu.:0.9275
##  Median :0.9710   Median :0.8605   Median :0.3723   Median :0.9420
##  Mean   :0.9660   Mean   :0.8558   Mean   :0.3822   Mean   :0.9390
##  3rd Qu.:0.9783   3rd Qu.:0.8862   3rd Qu.:0.4545   3rd Qu.:0.9565
##  Max.   :1.0000   Max.   :0.9213   Max.   :0.5909   Max.   :0.9855
##  Random Forest~ROC Random Forest~Sens Random Forest~Spec   Ridge~ROC
##  Min.   :0.8421    Min.   :0.2381     Min.   :0.9275     Min.   :0.7816
##  1st Qu.:0.8948    1st Qu.:0.4762     1st Qu.:0.9638     1st Qu.:0.8346
##  Median :0.9152    Median :0.5119     Median :0.9712     Median :0.8634
##  Mean   :0.9157    Mean   :0.5287     Mean   :0.9717     Mean   :0.8633
##  3rd Qu.:0.9383    3rd Qu.:0.5909     3rd Qu.:0.9784     3rd Qu.:0.8906
##  Max.   :0.9713    Max.   :0.8182     Max.   :0.9928     Max.   :0.9410
##    Ridge~Sens        Ridge~Spec
##  Min.   :0.0000   Min.   :0.9493
##  1st Qu.:0.1818   1st Qu.:0.9693
##  Median :0.2273   Median :0.9783
##  Mean   :0.2204   Mean   :0.9771
##  3rd Qu.:0.2727   3rd Qu.:0.9855
##  Max.   :0.4091   Max.   :1.0000
```



```
##    Resample          Elastic~ROC      Elastic~Sens     Elastic~Spec
##  Length:100         Min.   :0.7436   Min.   :0.5455   Min.   :0.6594
##  Class :character   1st Qu.:0.8444   1st Qu.:0.7619   1st Qu.:0.7536
##  Mode  :character   Median :0.8681   Median :0.8182   Median :0.7690
##                     Mean   :0.8645   Mean   :0.8124   Mean   :0.7747
##                     3rd Qu.:0.8894   3rd Qu.:0.8636   3rd Qu.:0.7975
##                     Max.   :0.9397   Max.   :1.0000   Max.   :0.8489
##    Lasso~ROC        Lasso~Sens       Lasso~Spec          LDA~ROC
##  Min.   :0.7581   Min.   :0.5909   Min.   :0.6884   Min.   :0.7619
##  1st Qu.:0.8363   1st Qu.:0.7532   1st Qu.:0.7477   1st Qu.:0.8397
##  Median :0.8673   Median :0.8182   Median :0.7754   Median :0.8706
##  Mean   :0.8655   Mean   :0.8129   Mean   :0.7742   Mean   :0.8654
##  3rd Qu.:0.8916   3rd Qu.:0.8636   3rd Qu.:0.8000   3rd Qu.:0.8916
##  Max.   :0.9455   Max.   :0.9545   Max.   :0.8696   Max.   :0.9382
##     LDA~Sens         LDA~Spec       Logistic~ROC     Logistic~Sens
##  Min.   :0.5909   Min.   :0.6884   Min.   :0.7522   Min.   :0.5714
##  1st Qu.:0.7727   1st Qu.:0.7391   1st Qu.:0.8412   1st Qu.:0.7619
##  Median :0.8182   Median :0.7690   Median :0.8637   Median :0.8182
##  Mean   :0.8201   Mean   :0.7688   Mean   :0.8650   Mean   :0.8129
##  3rd Qu.:0.8636   3rd Qu.:0.7971   3rd Qu.:0.8927   3rd Qu.:0.8636
##  Max.   :0.9545   Max.   :0.8696   Max.   :0.9393   Max.   :1.0000
##  Logistic~Spec       QDA~ROC          QDA~Sens         QDA~Spec
##  Min.   :0.6739   Min.   :0.7612   Min.   :0.5909   Min.   :0.6739
##  1st Qu.:0.7536   1st Qu.:0.8270   1st Qu.:0.7619   1st Qu.:0.7246
##  Median :0.7762   Median :0.8609   Median :0.8139   Median :0.7536
##  Mean   :0.7750   Mean   :0.8540   Mean   :0.8029   Mean   :0.7538
##  3rd Qu.:0.7989   3rd Qu.:0.8807   3rd Qu.:0.8636   3rd Qu.:0.7826
##  Max.   :0.8478   Max.   :0.9241   Max.   :1.0000   Max.   :0.8561
##  Random Forest~ROC Random Forest~Sens Random Forest~Spec   Ridge~ROC
##  Min.   :0.8326    Min.   :0.3636     Min.   :0.8986     Min.   :0.7332
##  1st Qu.:0.8982    1st Qu.:0.5455     1st Qu.:0.9420     1st Qu.:0.8389
##  Median :0.9240    Median :0.6364     Median :0.9565     Median :0.8673
##  Mean   :0.9209    Mean   :0.6261     Mean   :0.9517     Mean   :0.8632
##  3rd Qu.:0.9453    3rd Qu.:0.7143     3rd Qu.:0.9638     3rd Qu.:0.8881
##  Max.   :0.9802    Max.   :0.8571     Max.   :0.9928     Max.   :0.9381
##    Ridge~Sens        Ridge~Spec
##  Min.   :0.5000   Min.   :0.6739
##  1st Qu.:0.7273   1st Qu.:0.7477
##  Median :0.8182   Median :0.7754
##  Mean   :0.8015   Mean   :0.7724
##  3rd Qu.:0.8636   3rd Qu.:0.7971
##  Max.   :0.9545   Max.   :0.8849
```

Subset Selection Based on Variable Importance

A subset of predictors was selected based on their variable importance (varImp) scores, including alcohol, sulphates, volatile acidity, density, and citric acid. These predictors were deemed crucial for wine quality prediction based on their impact on model performance.

Performance Metrics with Subset of Predictors

Upon analyzing the performance metrics using this subset of predictors, it was observed that the ROC curves slightly performed worse compared to using all predictors. Models like Elastic Net, Lasso, and Logistic Regression showed a minor decrease in ROC curve values, indicating a slight reduction in overall predictive power when limited to the subset.

Sensitivity and Specificity Evaluation

Despite the decrease in ROC curve values, the subset of predictors exhibited improved sensitivity in most models, especially after upsampling. This improvement in sensitivity underscores the subset's effectiveness in capturing positive instances, which is crucial for identifying high-quality wines.

Comparison with Full Predictor Set

It's noteworthy that the subset of predictors did not significantly outperform or match the performance achieved with the full set of predictors. In fact, the subset technically performed worse in terms of ROC curve values, suggesting that the exclusion of certain predictors might have led to a loss of predictive information.

Upsampling Effect

When comparing the subset's performance with and without upsampling, sensitivity showed noticeable improvement post-upsampling, indicating that upsampling effectively addressed class imbalance issues and enhanced the models' ability to detect positive instances.

Conclusion on Subset Performance

While the subset of predictors based on varImp scores showed promise in improving sensitivity, it also resulted in a slight decrease in ROC curve values. This highlights the trade-off between sensitivity and overall predictive power when using a limited set of predictors. For optimal performance, considering all relevant predictors in conjunction with appropriate data balancing techniques like upsampling may yield more robust and accurate predictive models for wine quality prediction.