Final Project

The goal of the project is to model and understand the factors affecting wine quality. The data consists of:

1. Wine Quality (Response variable): A score between 0 to 10. 0 representing low and 10 representing high wine quality
2. Predictor Variables: fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, pH, sulphates, alcohol, and style

The data has been portioned into two (1) WineData.CSV, and (2) WineHoldoutData.csv. Use WineData.csv for model training, parameter tuning (if any), etc. WineHoldoutData.csv should only be used for evaluation of model performance. It should not be used in anyway in the model development process.

1. Develop SVM, Random Forests, and Boosting based regression model to predict wine quality. Perform hyper parameter tuning using 10-fold Cross Validation (CV). Summarize performance results and identify the best regression model. Report performance of best regression model on holdout data. 30 Points.
2. What are the assumptions made about wine quality data when using a regression model? Do you think it is justified to use a regression model on this data? 5 Points
3. Develop SVM, Random Forests, and Boosting based classification model to predict wine quality. Perform hyper parameter tuning using 10-fold Cross Validation (CV). Summarize performance results and identify the best classification model. Explain criterion used for selecting best classification model. Report performance of best classification model on holdout data. 30 Points
4. What information / detail about wine quality rating is lost when modeled as a classification problem? What kind of misclassification errors can this lead to? Based on this, suggest alternate supplemental metric that can be used in addition to standard misclassification rate. Document the "Misclassification Rate" and "Suggested Supplemental Misclassification Metric" for the three classification models on the holdout dataset. 15 Points
5. Analyze important features for regression and classification models. Are they consistent between regression and classification models? Across red and white wines? Across different techniques? What inferences you can make from the comparison? 10 Points
6. Try to use charts, tables and colors within R coding to present the results in a concise and impactful visualization. 5 Points
7. Write an "executive summary" summarizing results of regression and classification models with your interpretation. Summarize results on the holdout data. 5 Points