Alexander Overley

You are working with Cardinal Vineyard who plans to release 5 new wines to celebrate their 5th year as the leading wine makers of Kentucky. They have provided you with a dataset to help them focus their efforts of preferred combinations of acidity, alcohol content, and other metrics to determine which wines should be released. Their R&D team conducted blind tasting using wines from several different companies to build this data set provided. The final dataset will include the metrics from their proposed wines, a list of 20. Cardinal Vineyard wants your help to choose their best five wines. Based on your model, they will choose the 5 highest ranking wines from the final dataset that is to be provided.

Note: this company will add more data in the future to your dataset. They are conducting a second round of wine tastings to get more data for the model.

https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

User Stories

As a data scientist, I want to maximize variable feature importance in the model so that the best variables are used to determine the ratings.

DEFINED, IN PROGRESS, **COMPLETED**

As a data scientist, I want to utilize an ensemble tree method so that bias and variance can be reduced compared to simple decision trees.

DEFINED, IN PROGRESS, **COMPLETED**

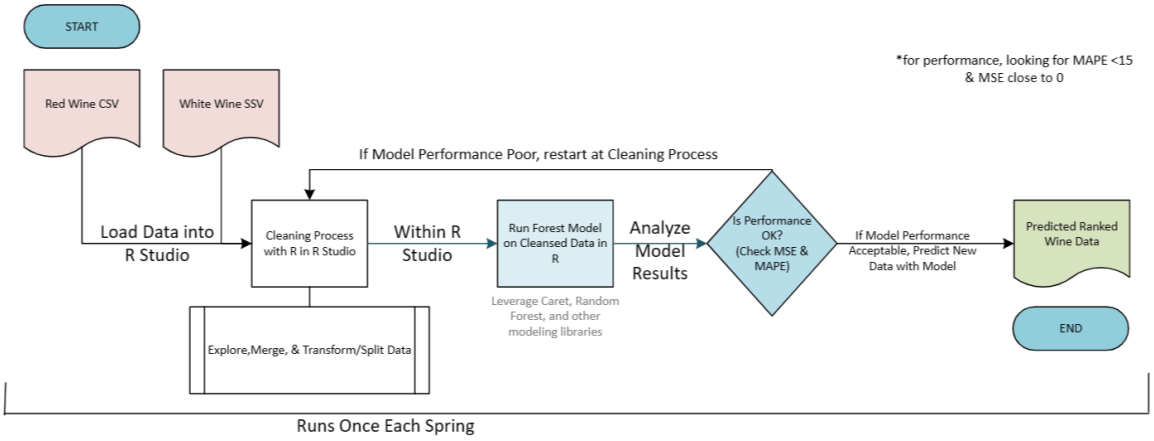
As a data analyst, I want to interpret the model results accurately so that the best wines can be chosen.

DEFINED, IN PROGRESS, **COMPLETED**

Best Wines

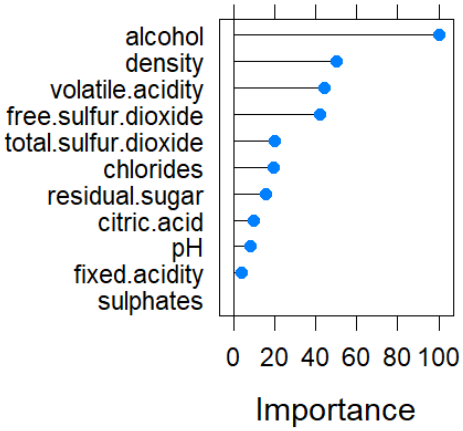
Based on the quality metric the best wines were index 11,13,16,17,20 (reference column 'index')

Pipeline

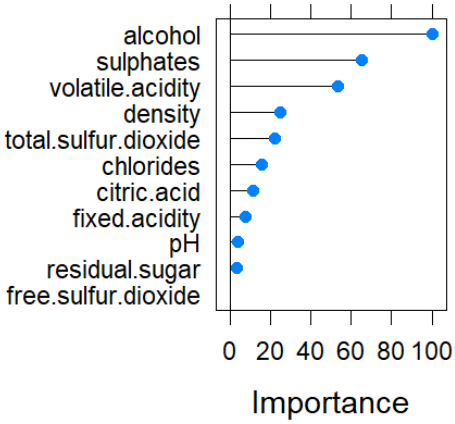


Alexander Overley

Notes



White Wines

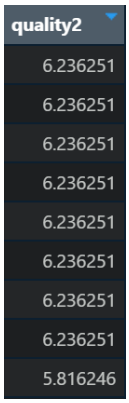


Red Wines

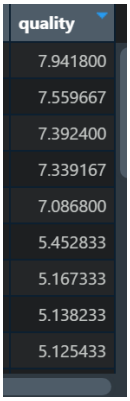
No variables were removed from the model. This was because of how the importance changes between red and white data and there were only 12 variables. It was decided that it was best to capture as much sensory information as possible to make the prediction.

For a basic tree model using caret's Rpart, the mean squared error (MSE) was 0.6 and the mean absolute percentage error (MAPE) was 10.94. The random forest MSE was 0.35 and MAPE 7.73 which is an improvement in the models performance. Accuracy in the model slightly increased and overall a random forest should perform better than a standard data tree because of bagging. The idea here is that the random forest is an aggregated result of many decision trees versus the result of one tree. This is seen in the final result comparison as the random forest definitively selected 5 top wines whereas the basic tree failed to do so and had a much more linear result.

Alexander Overley



Basic Tree



Random Forest