Predicting Wine quality

Introducion

The wine industry is an industry where the quality of wine is of much importance. With a variety of factors influencing the taste and appeal of wines, quantifying wine quality has become an interesting challenge. The dataset comprises different covariates, like chemical compositions and physical properties, which may influence wine quality. The objective of this study is to develop a model that can predict wine quality based on these covariates.

Methods:

I employed several data analysis techniques like Random Forest which is a bagging method that constructs multiple decision trees during training and outputs an average prediction of the individual trees for regression tasks. I also used XGBoost which is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It works by combining the predictions from multiple decision trees. Another technique I used was Bagging with Decision Trees which is an ensemble method that involves training decision trees on multiple bootstrapped subsets of the dataset and then averages the predictions. Finally I used Gradient Boosting Machine (GBM) which is a boosting technique that builds trees sequentially, each one correcting the errors of the previous one.

Data:

The datasets consist of red and white wine data, both inclusive of multiple variables related to the chemical composition and physical properties of the wines. Each wine entry is also tagged with a quality score. Initially, the datasets had missing values (NAs), which were promptly removed to ensure a smooth analysis.

Analyze:

For both red and white wines, Random Forest models were trained with 100 trees and with 3 variables tried at each split. The Root Mean Square Error (RMSE) for the red wine was 0.6090637, while for the white wine, it was 0.6773024.

The XGBoost models for both wine types were trained using parameters tailored for regression tasks (objective as reg:squarederror). The optimal number of boosting rounds for the red wine was determined using cross-validation and was found to be different from the white wine. RMSE values for red and white wines were 0.6134699 and 0.6818243, respectively.

Bagging with Decision Trees was used by training 100 decision trees on bootstrapped samples of the data, I obtained RMSE values of 0.6431839 for red wine and 0.7554135 for white wine.

Gradient Boosting Machine was trained with 5000 trees, a maximum depth of 6, and a shrinkage rate of 0.005. The resulting RMSE values were 0.6305585 for red wine and 0.7044085 for white wine.

Conclusion:

Considering the RMSE values, the Random Forest model was selected for the final predictions as it achieved the lowest RMSE for both types of wines. The model was then used to make predictions on the test dataset, which yielded a Kaggle score of 0.60875, just shy of the 18pt mark.

In the process of predicting wine quality based on various covariates, the Random Forest model emerged as the most accurate method, with an RMSE of 0.6090637 for red wine and 0.6773024 for white wine. I had a lot of challenges in this competition as I made extensive efforts to lower the RMSE, including parameter adjustments and trials with various algorithms.