ML Algorithm Selection

Introduction:

In this assignment, I am going to find the best ML algorithm to apply on wine quality dataset in order to get superior performance. Basically, I am going to select several ML algorithms, using them to predict the results, and comparing the results to find out which one is showing better performance. In this exercise, I used the regression learners including k-Nearest Neighbors (default parameters: k=5 and distance = 2), Linear regression (no specific parameter), Random Forest (default parameters: 100 trees and mtry=4), Support Vector Machine (default parameters: cost = 1.0 and epsilon = 0.1), and Gradient Boosting (default parameters: 100 rounds and max_depth=6). Finally, root mean square error was calculated for each model and compared to each other to provide more insight for ML algorithm selection process.

Dataset Description:

I am using the data from white wine which has 12 features and 4898 observations. Features include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality. I used all of the mentioned attributes to predict wine quality which has a numeric scale between 0 to 10. The dataset has no missing value. A summary of the dataset is presented in table 1.

Table 1. statistical summary of the white wine dataset

Feature	Min	Mean	Max	SD
Fixed acidity	3.80	6.85	14.20	0.84
Volatile acidity	0.08	0.27	1.10	0.10
Citric acid	0.00	0.33	1.66	0.12
Residual sugar	0.60	6.39	65.80	5.07
Chlorides	0.01	0.04	0.34	0.02
Free sulfur dioxide	2.00	35.31	289.0	17.00
Total sulfur dioxide	9.00	138.4	440.0	42.50
Density	0.98	0.99	1.04	0.003
pH	2.72	3.19	3.82	0.15
Sulphates	0.22	0.49	1.08	0.11
Alcohol	8.00	10.51	14.20	1.23
Quality	3	5.87	9	0.88

Experimental Setup:

I used R programming language and the libraries such as mlr3 (to implement ML tasks), mlr3learners (to train the models), mlr3viz (to visualize the results), readr (to read the data), caret (to split the data), and dplyr (to manage the data). I decided to fit the regression models to predict the wine quality, therefore, regression learners including k-Nearest Neighbors ("regr.kknn"), Linear regression ("regr.lm"), Random Forest ("regr.ranger"), Support Vector Machine ("regr.svm"), and Gradient Boosting ("regr.xgboost") were used in this exercise. To tackle this problem and select the best ML algorithm, I divided the data into two sets (training and validation). I applied 5-fold cross validation to the training dataset and calculated the RMSE for each model. It is noteworthy to say that CV technique is dividing the training dataset into subsets (training and testing) to fit the model. Afterward, I used the validation set to predict the results and again calculating the RMSE. Thus, we have RMSE for training (using 5-fold CV) and validation to compare the models together. In order to avoid bias and overfitting, we have to consider validation set. Calculation of the error just for train and test sets are not enough and can cause bias in our choice.

Results:

The results of the running model for both training (5-fold CV) and validations sets are presented in table 2.

Table 2. the results of the preciseness for different ML models using white wine data

Model	RMSE for 5-fold CV	RMSE for validation
k-Nearest Neighbors	0.70	0.68
Linear regression	0.75	0.75
Random Forest	0.62	0.61
Support Vector Machine	0.75	0.75
Gradient Boosting	0.66	0.63

We know the lower RMSE represents the better fit. Based on table 2, Random Forest showed better performance with RMSE = 0.62 in comparison with the other models. Both linear regression and SVM models were showing the worst performance with RMSE of 0.75. RMSE of 5-fold CV and validation are approximately equal for all of the models except KNN and GB which makes us more confident about performance of our ML models. The following plot (figure 1) is illustrating the prediction value against their true (observed) value using validation dataset for each regression model.

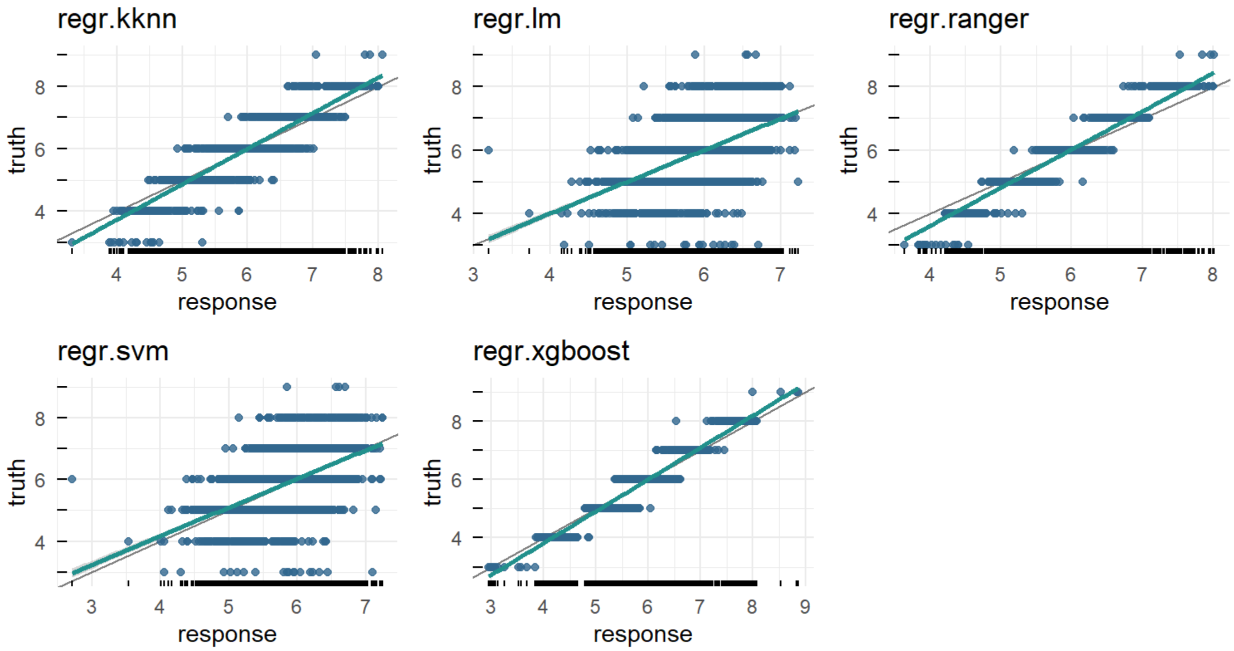

Figure 1. Scatterplot of the observed (truth) and predicted (response) values for quality of white wine using validation set.