Machine learning in biology

Assignment on Using multiple methods with the caret

Name: Anahi Quezada

Date: 10/30/2024

**Part 1 Classification**

Results of my models

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | AccuracySD |
| rpart | 0.9984848 | 0.004623160 |
| knn | 0.9612468 | 0.018429700 |
| C5.0 | 0.9969848 | 0.009226031 |
| svmRadial | 0.9697799 | 0.019439551 |
| Lvq | 0.9461254 | 0.024382124 |
| blackboost | 0.9919493 | 0.011024174 |
| earth | 0.9904116 | 0.016634798 |
| rf | 0.9984924 | 0.004600287 |
| AdaBoost.M1 | 0.9984924 | 0.004600287 |
| treebag | 0.9984848 | 0.004623160 |

The best-performing model I’ve applied for is C5.0 created by Ross Quinlan. This model predicts categorical outcomes using decision trees, with my data having columns with temperature, and relative humidity. I might say that the model is considering some patterns in the data. Since, going deep in my data the values of temperature and RH are recorded by hours. This model is also good with large amounts of data (9357 values) without using accuracy.

**Part 2 Regression**

Results of my models

|  |  |  |
| --- | --- | --- |
| Model | RMSE | RMSE SQ |
| lasso | 7.380562 | 0.04000922 |
| ctree | 7.576823 | 0.04154725 |
| gaussprLinear | 7.274348 | 0.04399727 |
| bayesglm | 7.273112 | 0.04439329 |
| kernelpls | 7.282112 | 0.04203029 |
| qrf | 7.853937 | 0.04099668 |
| bagEarth | 7.280482 | 0.05143551 |

After checking my results, I would say that bad Earth was the best-performed model. Looking to RMSE being the lowest value compared to the rest and RMSE sq having the highest value, this model indicates stability in executing my data. About this model, bagEarth, Specifically MARS (Multivariate Adaptive Regression Splines), this regression model presents a relationship with nonlinear segments or hinge functions. These relationships are between dependent and independent variables. In this case, with my air quality data where I used Benzene column “C6H6.GT” and Temperature “T”, Relative Humidity “RH” and “Absolute Humidity”. I would like to add that this dataset has non-linear interaction, and I think that’s why this model is the most suitable, reducing overfitting and creating stability for bagging.