Searching for the best Model. Wicked Science. Improving the model -- Due 12/07/2020

Raman Kannan, SPS Cuny, Fall 2020

PART-A

STEP#0: Pick any two classifiers of (SVM,Logistic,DecisionTree,NaiveBayes). Pick heart or ecoli dataset. Heart is simpler and ecoli compounds the problem as it is NOT a balanced dataset. From a grading perspective both carry the same weight.

STEP#1 For each classifier, Set a seed (43)

STEP#2 Do a 80/20 split and determine the Accuracy, AUC and as many metrics as returned by the Caret package (confusionMatrix) Call this the base_metric. Note down as best as you can development (engineering) cost as well as computing cost(elapsed time).

Start with the original dataset and set a seed (43). Then run a cross validation of 5 and 10 of the model on the training set. Determine the same set of metrics and compare the cv_metrics with the base_metric. Note down as best as you can development (engineering) cost as well as computing cost(elapsed time).

Start with the original dataset and set a seed (43) Then run a bootstrap of 200 resamples and compute the same set of metrics and for each of the two classifiers build a three column table for each experiment (base, bootstrap, cross-validated). Note down as best as you can development (engineering) cost as well as computing cost(elapsed time).

PART B

For the same dataset, set seed (43) split 80/20.

Using randomForest grow three different forests varuing the number of trees atleast three times. Start with seeding and fresh split for each forest. Note down as best as you can development (engineering) cost as well as computing cost(elapsed time) for each run. And compare these results with the experiment in Part A. Submit a pdf and executable script in python or R.

Part C

Include a summary of your findings. Which of the two methods bootstrap vs cv do you recommend to your customer? And why? Be elaborate. Including computing costs, engineering costs and model performance. Did you incorporate Pareto's maxim or the Razor and how did these two heuristics influence your decision?