Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer 1

The optimal value of alpha for ridge and lasso regression is 500. When I doubled the value of alpha for ridge and lasso, even ridge pushed the coefficients of 2 variables to zero and lasso immensely pushed many coefficients to zero making the model simpler. GrLivArea (Ground Floor Living Area) is the variable which is suggested by both lasso as well as ridge as the most important predictor.

Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer 2

I would choose lasso over ridge as it removes features by bringing down the values of its coefficients to zero. After doubling the values of alpha, I would go ahead with those values as it will make my model simpler as more coefficients has gone down to zero. The coefficient values have also made a significant loss by increasing alpha.

Question 3

After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer 3

They are:

1. GrLivArea

2. OverallQual

3. Neighbourhood

4. GarageCars

5. BasementExposureGood

Question 4

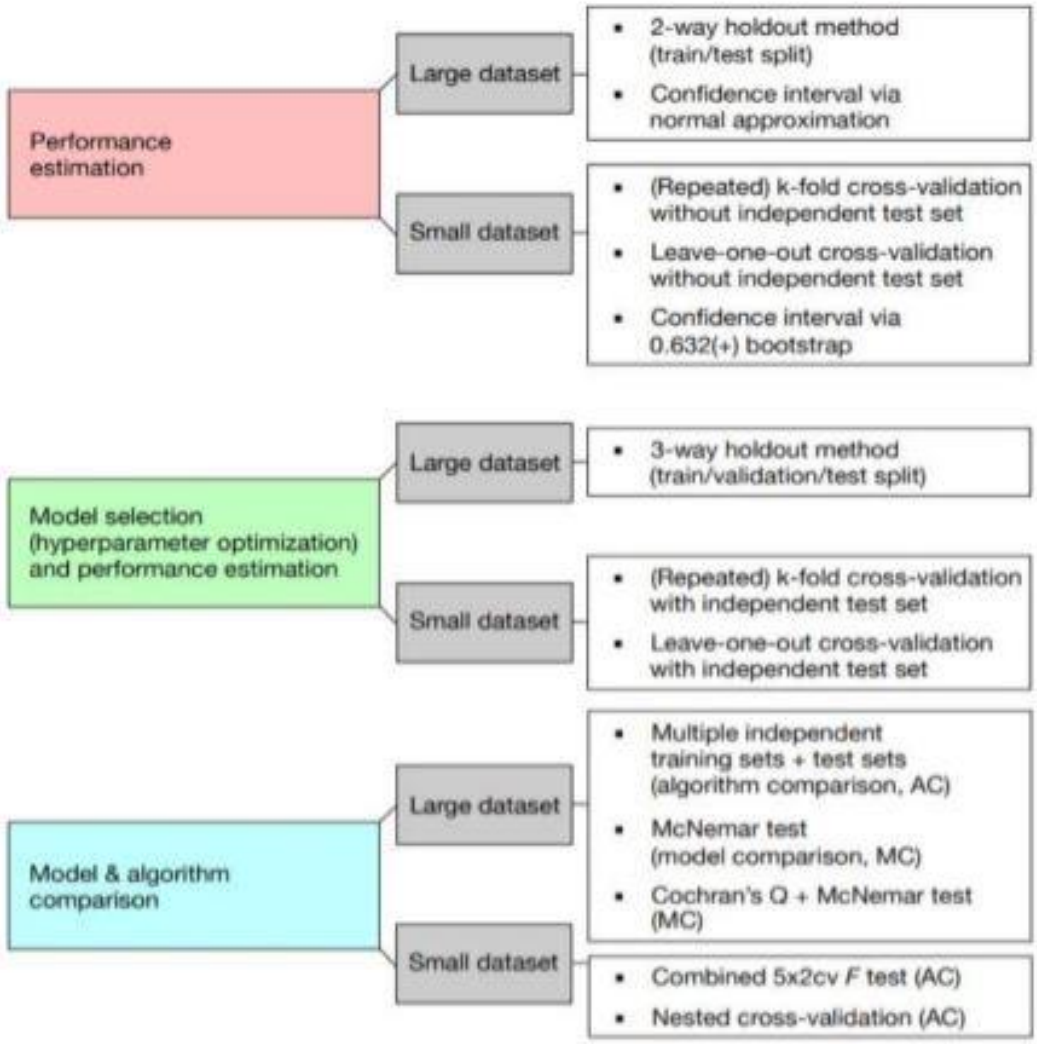
How can you make sure that a model is robust and generalizable? What are the implications of

the same for the accuracy of the model and why?

Answer 4

Since "a picture is worth a thousand words," I want to conclude this series on model evaluation, model selection, and algorithm selection with a diagram (Figure 23) that summarizes my personal recommendations based on the concepts and literature that was reviewed. It should be stressed that parametric tests for comparing model performances usually violate one or more independent assumptions (the models are not independent because the same training set was used, and the estimated generalization performances are not independent because the same test set was used.). In an ideal world, we would have access to the data generating distribution or at least an almost infinite pool of new data. However, in most practical applications, the size of the dataset is limited; hence, we can use one of the statistical tests discussed in this article as a heuristic to aid our decision making.

Figure 23: A recommended subset of techniques to be used to address different aspects of model evaluation in the context of small and large datasets. The abbreviation "MC" stands for "Model Comparison," and "AC" stands for "Algorithm Comparison," to distinguish these two tasks. Note that the recommendations I listed in the figure above are suggestions and depend on the problem at hand. For instance, large test datasets (where "large" is relative but might refer to thousands or millions of data records), can provide reliable estimates of the generalization performance, whereas using a single training and test set when only a few data records are available can be problematic for several reasons discussed throughout Section 2 and Section 3. If the dataset is very small, it might not be feasible to set aside data for testing, and in such cases, we can use k-fold cross-validation with a large k or Leave-oneout cross-validation as a workaround for evaluating the generalization performance. However, using these procedures, we have to bear in mind that we then do not compare between models but different algorithms that produce different models on the training folds.



Nonetheless, the average performance over the different test folds can serve as an estimate for the generalization performance (Section 3) discussed the various implications for the bias and the variance of this estimate as a function of the number of folds).

For model comparisons, we usually do not have multiple independent test sets to evaluate the models on, so we can again resort to cross-validation procedures such as k-fold cross-validation, the 5x2cv method, or nested cross-validation. As Gael Varoquaux [Varoquaux, 2017] writes: Cross-validation is not a silver bullet. However, it is the best tool available, because it is the only non-parametric method to test for model generalization.