Lecture 4 Customer Analytics – Subset Selection, LASSO, Decision Trees, Random Forests

Questions received via email

1. [Kevin];
   (1) In slide 11 & 12 you talk about the estimates becoming non-zero when lowering lambda. Am I correct that the variables that are the most influential, or 'important', in predicting y will enter the model first?

   (2) Can we add interaction terms in these approaches? Should we then just create new variables that take the product of the two variables?

   (3) A little out of context maybe, but:
   Since we look at a lot of variables without having a theory a priori, what is the relation of the stepwise regression method to p-hacking (data-dredging)? Can we consider that this is not a problem since we use OOS data to validate our findings?

2. [Konstantinos];
   (1) In decision trees, we apply k-fold CV to define the proper number of nodes. In other words, it means that we apply it to find the proper threshold (to stop the splitting)?

   (2) Does it make sense to say that models based on bagging will tend to overfit compared to the random forest ones?

3. [Noa];
   (1) When choosing a lambda, we choose one such that the deviance is minimized. Are we talking about OOS deviance here?

   (2) On slide 22, we compute the Gini index. Could you explain the computation behind potential split 2 on this slide?

   (3) Regarding random forest (slide 35), we pick recency as our root node. As we are using random sampling, does this then also imply that we can choose recency in another internal node again? Is it like bagging (so with replacement) or not?

4. [Mieke]
   (1) On slide 10, what do you mean with ‘the shape of the penalty function means that a lot of betas will be exactly zero’?

   (2) On slide 34, I do not understand why this extension of the random forest makes the observations uncorrelated. Could you please clarify this?

(3) On slide 35, how do you determine these tresholds at every split in the three? For example, R > 1 and R <= 1