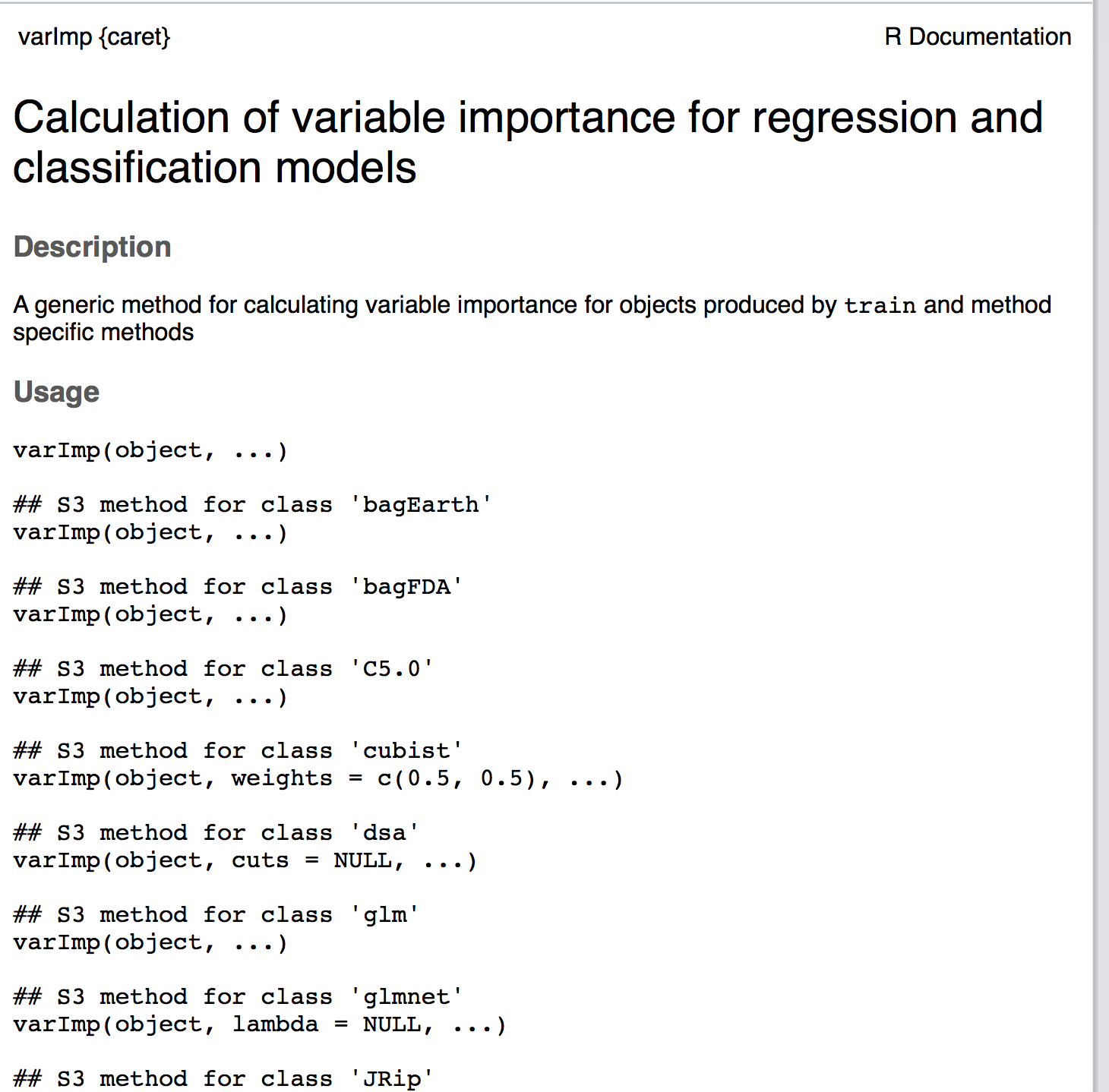
1. I am little bit confused about why we prefer to use Gini and Entropy, not misclassification rate to check the cutoff and variable for each node?

🡺 We just believe these two are better ways to find impurity

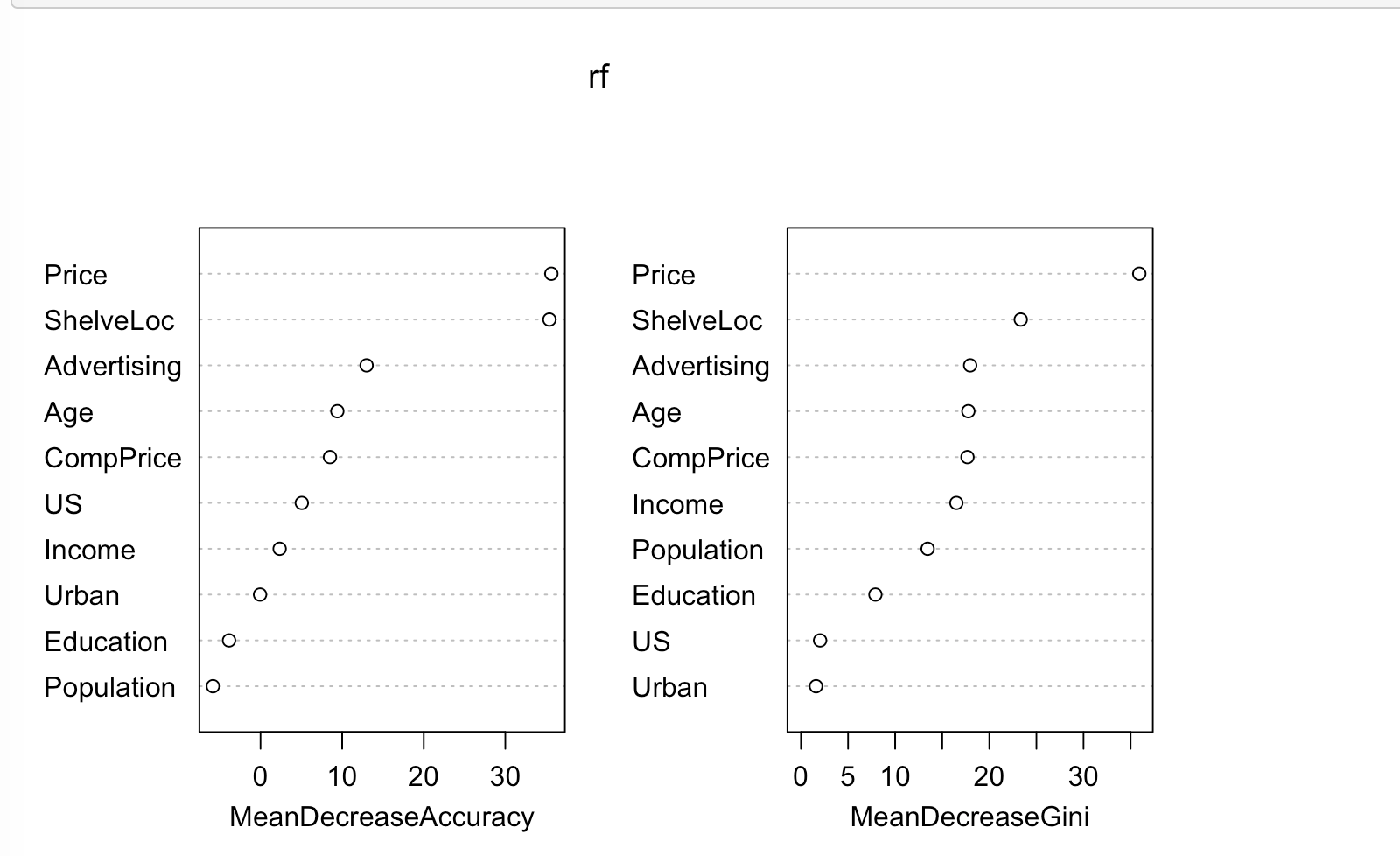
1. Do we do EDA on both training and testing? 🡺 you can clean the data & but scaling things (sd and mean should be from training) & discretizing and dummifying for both are fine.
2. So, I guess we are using CV to do cost-complexity pruning. Right? 🡺 You can do it.
3. Do we find outliers for categorical variables? 🡺 just don’t do it.
4. So our goal is reduce the test prediction error. Do we have to reduce error only with the given methods, or are we allowed to use any method to achieve the goal? 🡺 Its more like you understand the methods. But, definitely, you can use more methods.
5. Why do we really care about OOB error rate if this is basically the error rate from training set…?? 🡺 rough estimate/expectation for prediction error.
6. So, basically, do we dummify all the categorical variables and discretize all the continuous variables? So, at the end, there is no continuous variables??? 🡺 we dummify all the categoricals, and it makes sense.

But for continuous variables, we don’t just discretize all of them. But “native.countires” makes sense to discretize since most of them are “united states”

1. When are we going to get invited to box folder? 🡺 this week
2. When do R coding, do we separate function and scripts like what 133 did? 🡺 can leave it as rmd file.
3. More information of how to write and how much to write a report?
4. Is there any restriction in the project? For example, using embedded functions or so..? 🡺 No!!!
5. Do we need to make slides? 🡺 No.
6. How are you going to grade this? Is there any rubric? 🡺 description, typo, visual
7. So, I guess we do not need to normalize/scale for any tree methods because it might be harder to interpret? 🡺 Not scale and explain why you did not scale it…
8. What do you mean by “variable importance statistics”? Are you saying that important variables? Can you give an example? 🡺 varImp from caret library.



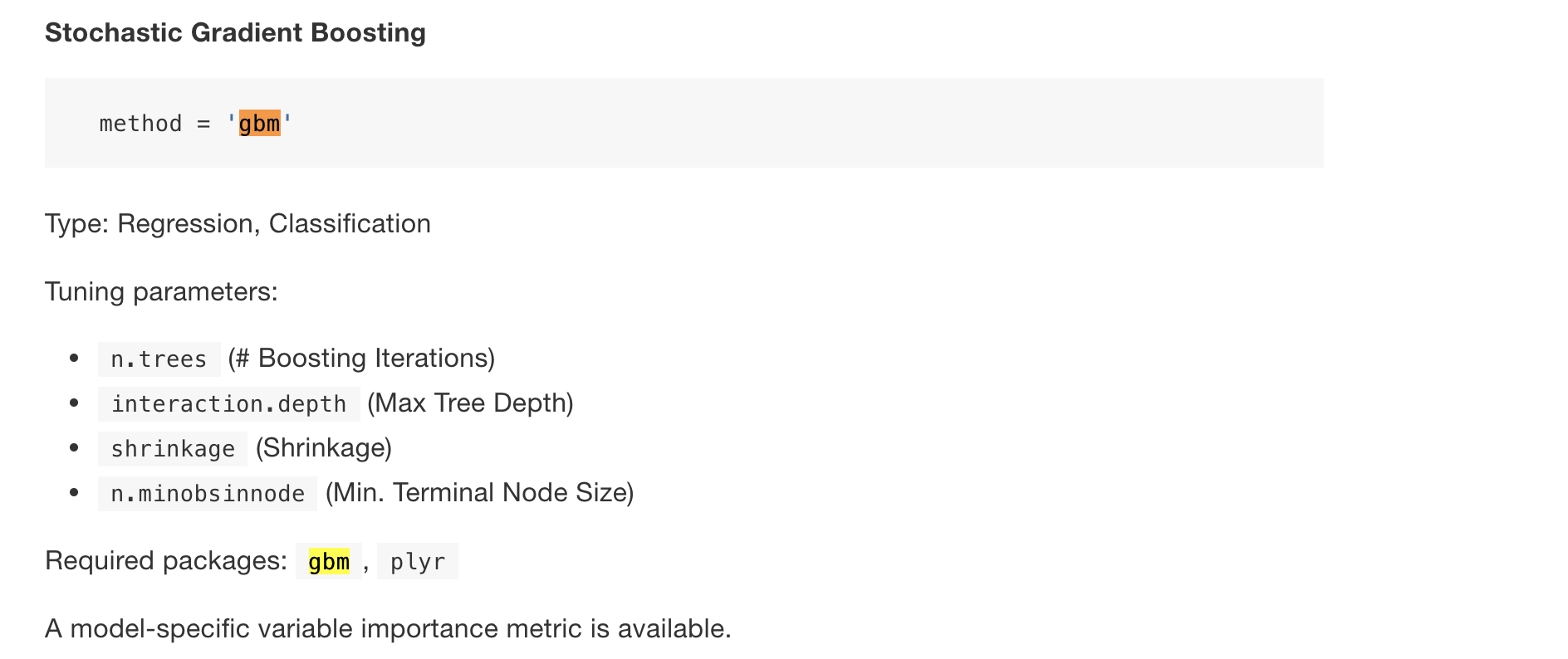
1. How to interpret variable importance with the tree output?



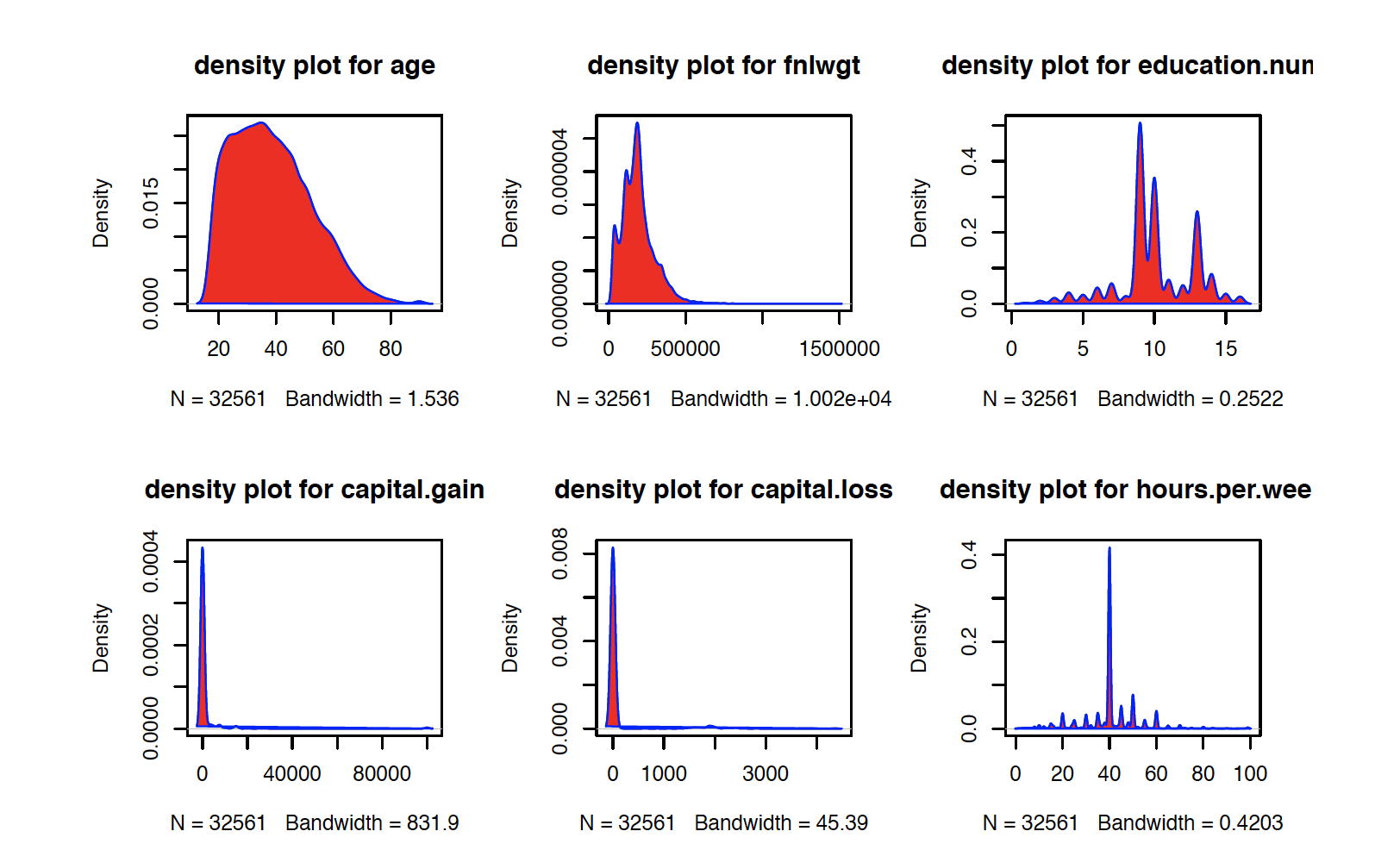
so price and shelveloc are the most important variables as the meandecrease are huge if they are removed…

1. Can you please check whether tuning parameters for each method I have are comprehensive?:

* Classification: alpha for cost complexity, classification for splitting tree
* Bagged: number of trees (using tuning parameter alpha for cost complexity), depth of the tree/minimum size of nodes
* Random forest: number of predictors considered, number of trees, depth of the tree/minimum size of nodes
* Boosted: number of splits/interaction depth, shrinkage parameter, number of tree



1. Although we say there are a set of number of tuning parameters, is it possible for us to tune more? For example, we usually only tune alpha for classification, but can we actually tune something else such as minimum number of observations in terminal node? 🡺 yes. You can tune anything you want!!!
2. Since many of the methods take forever, we try to simplify the model as much as we can, like making n.tree be smaller, etc. Do you think this is fine for us to simplify parameters if error rate is almost the same? ==> if your goal is for prediction error rate, making complex is better. But, if the goal is making it interpretable, making simple is better. So, make the reasonable tuning, and say why you tune upto there. Make a reasonable explanation.
3. So, when you said “model selection,” does it mean that which are the best out of classification tree, bagged tree, and random forest? 🡺 1. Pick from the parameters from the same model 2. Pick from the different models
4. Need to check my missing value imputation…. 🡺 Stil make sense, but you need to double check, whether imputed vluaes still make sense…. ?
5. Can we just take out “99999” for capital gain variable as outliers? How do we know they are outliers? 🡺 either remove or do another missForest to find missing values.
6. Do we need to remove mathematical outliers, or can we just leave them if we think they ? 🡺 argue it then it will be fine…
7. Do we remove linearly dependent variables (education and education.num)? 🡺 Yes! Sure. Find correlation of each variable, and drop the most correlated ones
8. Do you think it make sense for us to do regression tree, SVM, LDA/QDA/MDA/logistics, boosted, cluster, PCA for our project? 🡺 pca might be good for EDA
9. And, I am not sure whether I am understanding fnlwgt variable correctly. Is it possible to have huge number for fnlwgt? What is the boundary for fnlwgt? 🡺 not sure. You can ignore or keep them, it might be useful. But, feel like this should not be your predictors… (make sure that our data is subsample of the entire data)
10. Do we need to dummify the discretized variables as well? 🡺 Yes!
11. Do you think it makes sense to discretize any of them? 🡺 No…



1. Need to check my discretization for categorical variables…
2. Can discretizing also mean combining several categories into one? 🡺 Yes~!
3. When we do dummify and discretize, we apply them both on training and testing sets. Right? 🡺 Yes! (cuz it is data preprocessing…)
4. What do you mean by “Association between each predictor and the response”in EDA? 🡺 plotting and roughly see how predictors related to income.
5. When we dummify, since there is no intercept we need to do one hot encoding, then does it mean that we always have one column less than the level of the variable. Then, will it be hard to interpret it after we do tree as there are so many dummified variables?
6. How many intervals to make when discretizing? Also, is it make sense for us to combine training and testing “only for this part” to have the same intervals for discretizing?
7. If my trained function (fixed shrinkage) performs worse than the other model with higher shrinkage in both testing set AUC, (~ I guess it is possible because our model has not really tuned shrinkage) and

worse than the other model with higher tree numbers in only for training set. What should I choose? 🡺 My train function is giving optimal value with the highest accuracy as I said metric = “Accuracy”… So, its possible for us to have lower AUC.

Seems like higher shrinkage parameter gives you better testing AUC, but it has higher variability, so can we say it and go with model with higher AUC but lower shrinkage parameter?

So, if the model A has lower training set AUC, but higher testing set AUC than the model B, then I should choose model B. Right? 🡺 Yes. Testing set AUC is more important.

If so, we basically choose the model with low test set AUC, then why do we even get train set AUC?

Is it possible for the model with smaller AUC better? 🡺 Right!!!

1. Do we need to use accuracy rate or AUC/ROC to decide the best model? How are they different? Don’t we need to use accuracy rate for every model selection with the certain threshold, because AUC/ROC is used when we think one type of false rate is worse than the other type of false rate? (because our goal is not trying to decrease certain type of error rate, but try to improve prediction accuracy)

🡺 if goal is prediction: accuracy rate & if there is any class imbalance (80 v.s. 20) in entire data set: AUC/ROC

If so, what threshold do we need to use for accuracy rate? 🡺 tune in training set it to get highest accuracy rate.

1. Need to check the way how I got ROC curve, as ROC function for gbm does not work… Also, how many trees to use in predict function when I try to collect posterior? 🡺 either you decide based on the error rate plot or just follow the number of trees we used for train….
2. Also, I need to check my model selection work, especially, how the number of trees I said. 🡺 same as above….
3. Last time you said interaction depth is learning rate. But, isn’t learning rate a shrinkage parameter? 🡺 right!
4. Is it even possible to draw boosted tree plot? 🡺 No! Not possible to draw a lot of trees here….
5. The instruction says “report the training accuracy rate”. But doesn’t it have to be “testing accuracy rate”? Or, do we have to get both of them? 🡺 both!
6. I am confused with the concept of fit a tree to the residual from the model in boosting tree. What do you mean by residual here? 🡺 the missed data….
7. Need to check boosted plots for when shrinkage is changed to 0.1 or 0.2… 🡺 slightly over-fitting…

* Choose number of trees based on test error rate (say my threshold is reasonable) and then find the threshold giving the best prediction accuracy using cross validation.

1. The instruction says “report the training accuracy rate”. But doesn’t it have to be “testing accuracy rate”? Or, do we have to get both of them? 🡺 both!
2. Do we need to set.seed for all of decision tree, boosted, bagging, random forest to get the same answers?
3. Is it possible for the trained function to have higher testing set error than the other functions?=> Yes!
4. Professor said when the threshold is 0.5, the misclassification rate will be the lowest… But why not the case for ours..? 🡺 threshold = 0.5 then we are weighting two errors rate the same. (sort of right!!!)
5. If train function has lower accuracy in testing set, how do we actually get so called “optimal” function in testing set? And, what is the point of using train function in training set if other functions can perform better in testing set? 🡺 We just expect to do well in testing set if we do in training set.
6. Do you want us to turn in as just one Rmd or separate Rmd based on methods? 🡺 upto you.
7. Is my way getting the good threshold right, although I am not getting it with Cross validation? (compare training set error with threshold testing on entire training set!) 🡺 Yes! If cv takes a lot of time.
8. Can we select one model for best AUC and one model for best testing prediction accuracy (or those two might be the same model), and conclude both of them as my best model? 🡺 Yes! Right! But make it clear!!!!
9. So, for boosted and decision tree methods, we plug in different number into parameters and get accuracy rate and ROC, etc. And, we do this several times to see how different parameters change variance important statistics and accuracy rate, etc. And, then we use train function to get the best model/parameters. And, we compare all of them together into ROC or accuracy rate to prove our best one is really the best. So we are kind of doing the inner model selection. Is this correct way to do this project? 🡺 Yes
10. For ROC curve, does it even make sense to get it predicting on train set? So, in our project, do we do both or only on testing set?
11. How to use latex to print the confusion matrix?
12. How to adjust pictures/captions locations in R markdown latex when we use it for report?
13. Although our model is trained with 150 trees, am I able to include more number of trees in predict function?
14. When we dummify categorical variables, do we take out one column since it can be linearly dependent?
15. Any hints how to do well in cluster?
16. I guess relative influence of variables and variance importance actually mean the same thing..??
17. For boosted, I got training error and testing error rates are so similar. Are we not supposed to get something like this??
18. I found that when you predict for boosting using gbm function, you can actually include response variable and similar answer with the one without including it. But, still, we should not in include response when we predict. Right?
19. Are we allowed to use different types of boosting, bagging, random forest? (<http://topepo.github.io/caret/train-models-by-tag.html#boosting)>
20. APM pg 381 says that when there is a missing value, we need to weight it differently. Can you explain it to me in detail how we should deal with missing values…?

* Structurize all the image formats with latex.
* Grammar/Format check (space, decimal points, etc) for reports