## ***Goal: Based on 15 variables on individual samples, perform prediction to determine whether a person makes over 50K a year!***

## 

## **Preprocessing and Exploratory Data Analysis**

<https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/tutorial-random-forest-parameter-tuning-r/tutorial/>

<https://mathematicaforprediction.wordpress.com/2014/03/30/classification-and-association-rules-for-census-income-data/>

<http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names>

<https://www.analyticsvidhya.com/blog/2016/08/practicing-machine-learning-techniques-in-r-with-mlr-package/>

<https://stats.stackexchange.com/questions/49270/imputation-with-random-forests>

<https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/>

**Start your analysis cycle with an exploratory phase so you get to know and understand the data set. Below is a (non-comprehensive) list of (optional) considerations to keep in mind:**

* **Handling missing values**

**1. Find missing value or NA -> use sum(is.na) or use the way I did in pbhlth 245 project... & variable == "" to find empty place -> Think of how to fix these issues.... (i.e. simple randomly assign, plug in the most frequent values, median of the rest data, or use predict function...)**

* **Take out the variables that are not interpretable**

**4. Find out continuous variables, and transform them into numeric if necessary.**

**5. Find out categorical variables, and transform them into factor if necessary.**

* **Handling outliers with str and summary to find out data structures... (fix if necessary)**

**7. Boxplot**

**10. density estimators plot**

* **Visualizing distributions**
* **Changing scales**
* **Binning (i.e. discretizing)**

**-> easier to divide up or interpret when doing tree.**

* **Converting to (dummy) indicators -> need to explain why I not do “grouped categories” but “independent categories” on pg 373 APM**

**APM pg 374 says “for classification trees using CART, there is no practical difference in predictive performance when using grouped categories or independent categories predictors” (1-hot encoding)**

1. **Easier to interpret**
2. **As APM pg 377 says, independent category predictors provide valuable interpretation about relationship between predictors and response.**

* **Summary statistics**
* **Association between each predictor and the response**

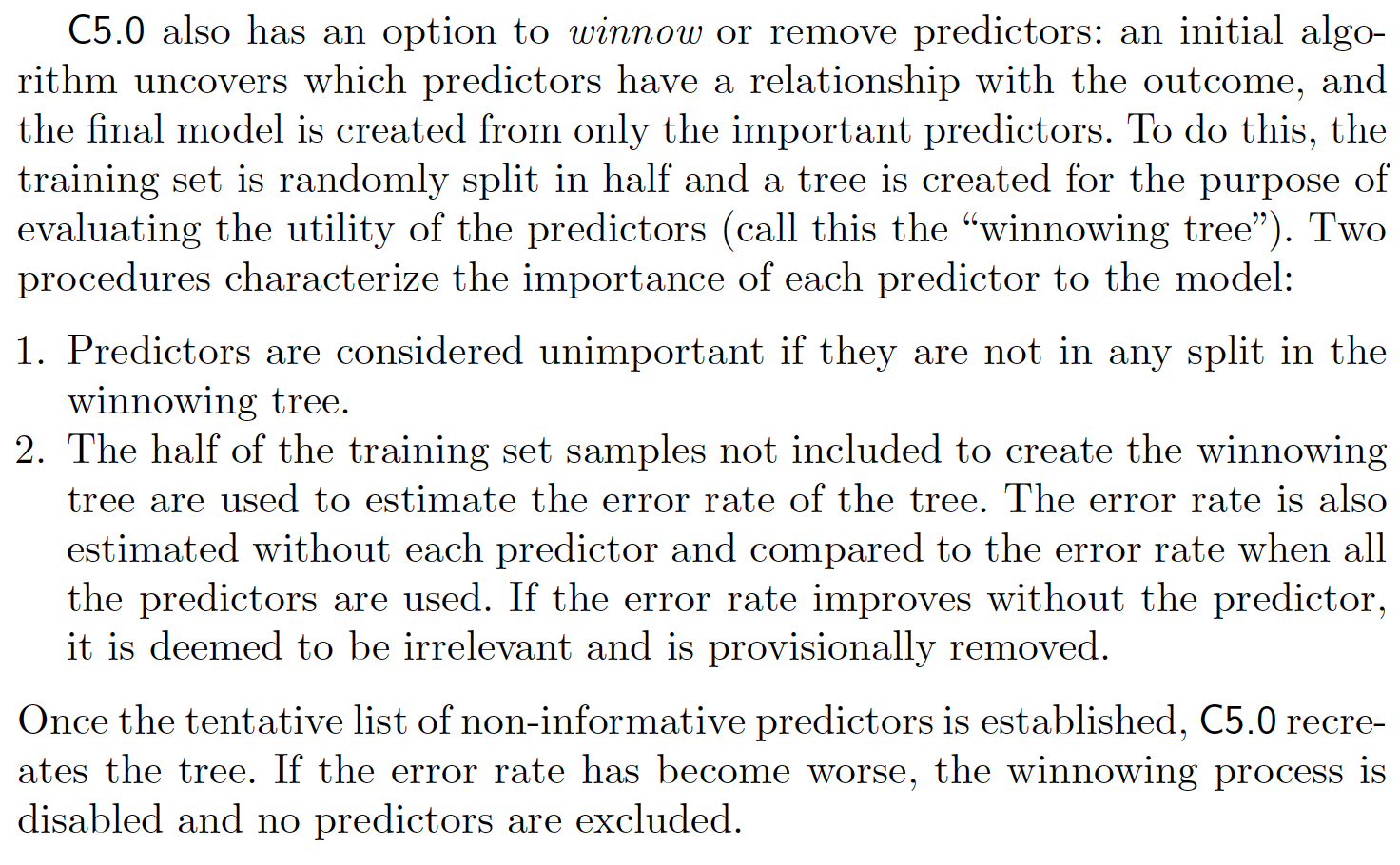
**6. scatter plot**

**8. pairs plot with smoothing lines**

**9. co-plot**

**11. Cook's distance & leverage & residual plots**

**Pg 398 of APM**

****

### **Build a Regression Tree**

* **Cost complexity pruning/weakest link pruning with tuning parameter alpha (choose alpha with K-fold CV)**
* **Follow algorithm on pg 309**
* **Draw graph on pg 311**

### **Build a Classification Tree - lab12**

<https://www.analyticsvidhya.com/blog/2016/08/practicing-machine-learning-techniques-in-r-with-mlr-package/>

<https://sadanand-singh.github.io/posts/treebasedmodels/>

<https://github.com/ucb-stat154/stat154-fall-2017/blob/master/papers/Intro-to-rpart.md>

<https://mathematicaforprediction.wordpress.com/2014/03/30/classification-and-association-rules-for-census-income-data/>

**<Library: tree, C50, partykit, rpart, RWeka, tree, party>**

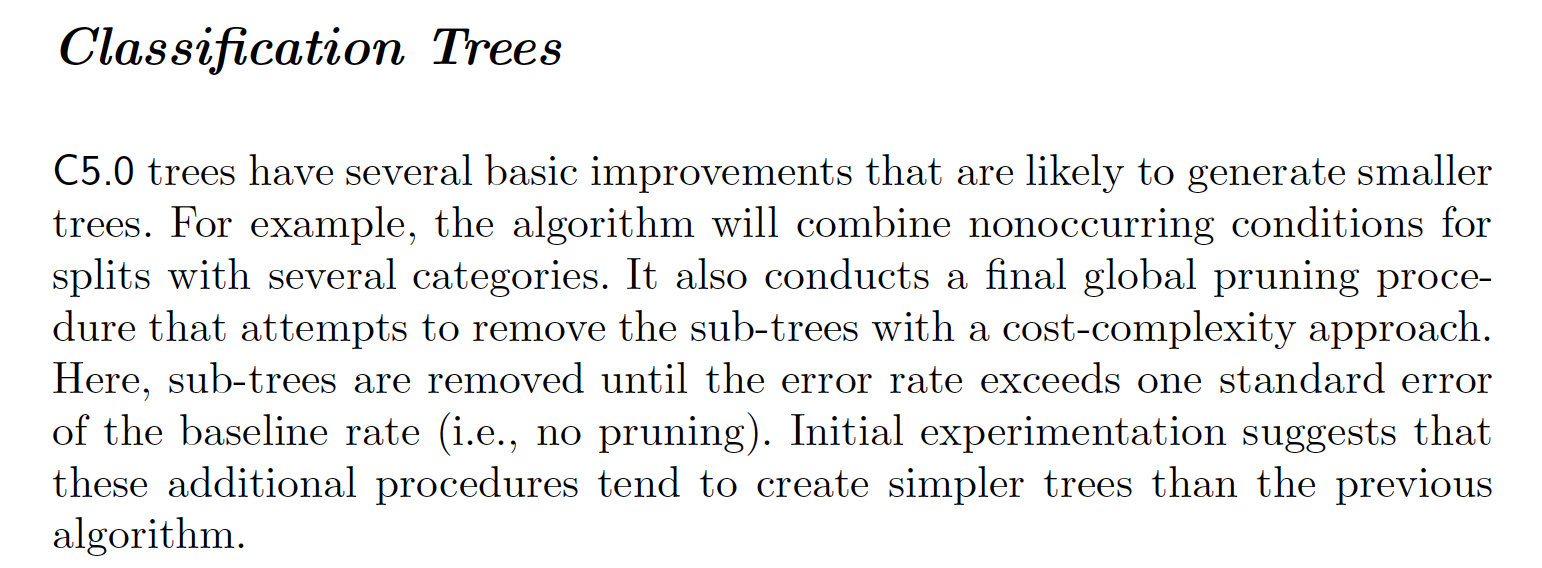
**<Be careful: when predictor is categorical (pg 372 APM), >**

* **Draw figure 8.2 in pg 305**
* **Fit a classification tree .**
* **Make plots (pg 313) and describe the steps you took to justify choosing optimal tuning parameters. (number of trees - Cost complexity pruning/weakest link pruning with tuning parameter alpha (choose alpha with K-fold CV), )**
* **Report your 7 important features, with their variable importance statistics.**
* **Report the training accuracy rate.**
* **Plot the ROC curve, and report its area under the curve (AUC) statistic. (maybe with confusion matrix)**
* **Do class prediction and class proportion among training observations (pg 311)**
* **Talk about Gini index and Cross-entropy for all split points (node purity is important -> pg 314) instead of minimizing misclassification error (APM pg 370)**

**-> First, order based on proportion of samples in selected class and and Gini and draw plot like pg 374 APM.**

* **Try to make Gini index table like the one in APM pg 371**
* **Pruning…. -> smaller tree with fewer splits lead to lower var and better interpretation at cost of a little bias ~~> select subtree that leads to lowest test error rate using CV…**
* **Compare test error rates with other classification methods…**

**Pg395 of APM**

****

### **Build a Bagged Tree**

**<Library: randomForest, ipred>**

* **Train a Random Forest classifier**
* **Make plots and describe the steps you took to justify choosing optimal tuning parameters. (# of trees)**
* **Report your 7 important features, with their variable importance statistics.**
* **Predict with “majority vote” (pg 317) & out-of-bag(OOB) classification error (318) without doing CV.**
* **Report the training accuracy rate.**
* **Plot the ROC curve, and report its area under the curve (AUC) statistic. (maybe with confusion matrix)**
* **Obtain an overall summary of importance of each predictor by adding total Gini index and averaging all trees (pg 319)**

### **Build a Random Forest - lab12**

<https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/tutorial-random-forest-parameter-tuning-r/tutorial/>

<http://www.rmdk.ca/boosting_forests_bagging.html>

<https://www.analyticsvidhya.com/blog/2016/08/practicing-machine-learning-techniques-in-r-with-mlr-package/>

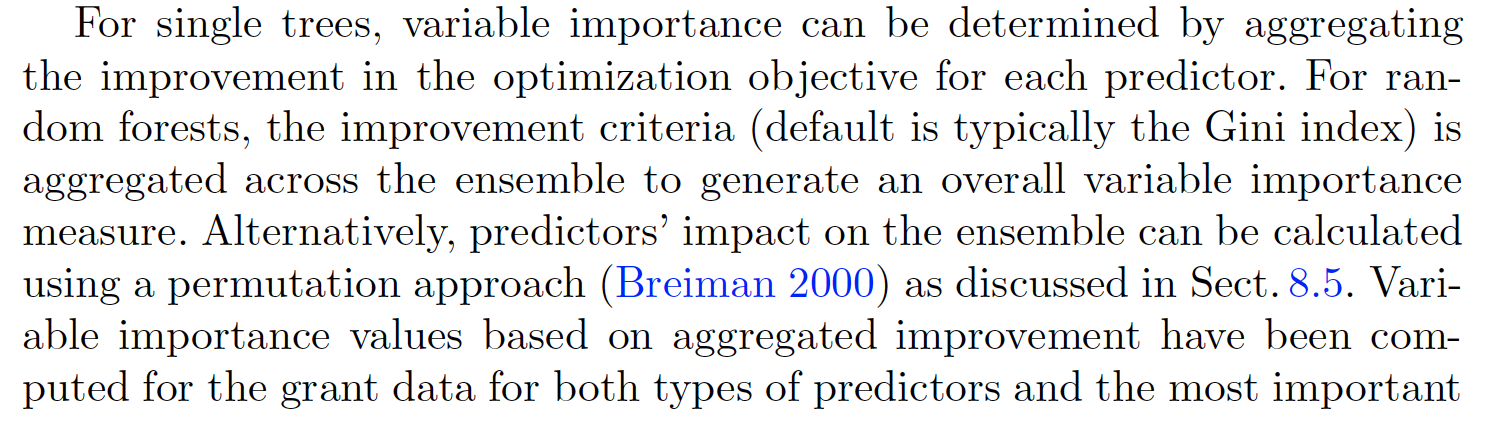
**<Library: randomForest, ranger, parallelForest, RRF>**

* **Train a Random Forest classifier**
* **Make plots and describe the steps you took to justify choosing optimal tuning parameters. (the number of predictors considered at each split - ‘mtry’, number of trees to build before majority vote - ‘ntree’, the depth of the tree - ‘nodesize’, minimum size of parent and leaf nodes)**

**\*While the type of tree changes in the algorithm, # of randomly selected predictors to choose from at each split is the same!**

* **OOB error rate**
* **Report your 7 important features, with their variable importance statistics.**
* **Report the training accuracy rate.**
* **Plot the ROC curve, and report its area under the curve (AUC) statistic. (maybe with confusion matrix)**
* **Obtain an overall summary of importance of each predictor by adding total Gini index and averaging all trees - importance() and varImPlot()**

**Pg 388 of APM**

****

**Build a Boosted Tree - lab12**

[**http://www.rmdk.ca/boosting\_forests\_bagging.html**](http://www.rmdk.ca/boosting_forests_bagging.html)

[**https://www.analyticsvidhya.com/blog/2016/08/practicing-machine-learning-techniques-in-r-with-mlr-package/**](https://www.analyticsvidhya.com/blog/2016/08/practicing-machine-learning-techniques-in-r-with-mlr-package/)

[**https://stats.stackexchange.com/questions/103495/how-to-find-optimal-values-for-the-tuning-parameters-in-boosting-trees**](https://stats.stackexchange.com/questions/103495/how-to-find-optimal-values-for-the-tuning-parameters-in-boosting-trees)

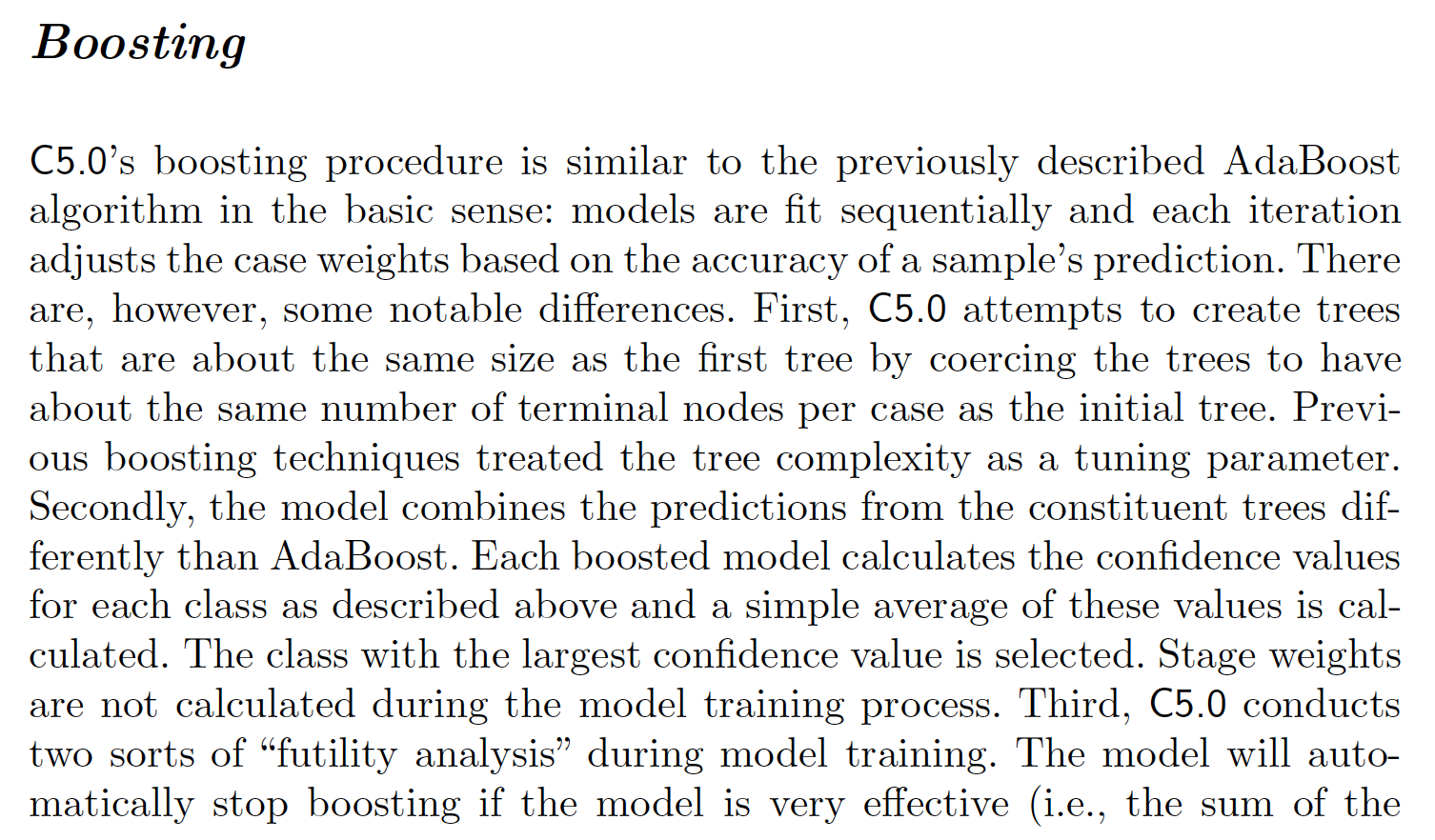
[**http://www.listendata.com/2015/07/gbm-boosted-models-tuning-parameters.html**](http://www.listendata.com/2015/07/gbm-boosted-models-tuning-parameters.html)

[**https://www.analyticsvidhya.com/blog/2016/08/practicing-machine-learning-techniques-in-r-with-mlr-package/**](https://www.analyticsvidhya.com/blog/2016/08/practicing-machine-learning-techniques-in-r-with-mlr-package/)

**<Library: gbm>**

* **Make plots and describe the steps you took to justify choosing optimal tuning parameters (interaction.depth [number of splits], shrinkage parameter, number of trees... (322))**
* **Use algorithm in pg 323**

**Pg 396 of APM**

****

### **Model Selection**

* **Draw Figure 8.8 on pg 318 to compare all methods...**
* **Validate your best supervised classifier on the test set.**
* **Compute the confusion matrix.**
* **Using the class "over 50K a year" as the positive event, calculate the *Sensitivity* or *True Positive Rate* (TPR), and the *Specificity* or *True Negative Rate* (TNR).**
* **Plot the ROC curves of all the classifiers. & AUC (misclassification error rate curves)**

**1. Logistic regression with a full basic model, with survival status being response (make sure to include family = "binomial") onto the training data set**

**2. Perform different model selection methods: adjusted $R^2$, AIC, BIC, Mallow's Cp, on to the full model -> use either bestglm**

**3. Include interaction term or remove some variables**

**4. Perform different model selection methods: adjusted $R^2$, AIC, BIC, Mallow's Cp, on to the new model -> use either bestglm**

**5. Get the Best model/model selection methods by comparing misclassification rate with confusion matrix or ROC curve inside of the training set**

**6. See how residual deviance gets better than the full model**

### **Build a SVM - lab 11**

<https://www.analyticsvidhya.com/blog/2016/08/practicing-machine-learning-techniques-in-r-with-mlr-package/>

* **Do the similar thing as lab**

### **Build a Cluster**

* **Do Mclust()**
* **Apply kmeans to the data with the best number of clusters found with Mclust**
* **Do hierarchial clustering with different linkages**

### **Build a PCA**

* **Correlation circle**

### **Build LDA, QDA, MDA, KNN**

<https://www.analyticsvidhya.com/blog/2016/08/practicing-machine-learning-techniques-in-r-with-mlr-package/>

* **Report misclassification rate with different models**
* **Tune K in KNN**

[**https://mathematicaforprediction.wordpress.com/2014/03/30/classification-and-association-rules-for-census-income-data/**](https://mathematicaforprediction.wordpress.com/2014/03/30/classification-and-association-rules-for-census-income-data/)

[**https://www.quora.com/What-is-the-meaning-of-capital-gain-capital-loss-and-fnlwgt-in-adult-dataset-from-UCI**](https://www.quora.com/What-is-the-meaning-of-capital-gain-capital-loss-and-fnlwgt-in-adult-dataset-from-UCI)

[**https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/**](https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/)

[**https://www.kaggle.com/general/4092**](https://www.kaggle.com/general/4092)