Statistical Learning Methods, summer semester 2017/18 – Practical exam, groups 100 & 101

**During exam you can use Internet, codes, books or notes but communication with each other (and other people) is forbidden.**

**Ex. 1 (12 p)**

Investment fund XYZ considers buying personalized software which can be used for forecasting of stock prices. Price of such software amounts to $35.000. If the company decides to buy this software, it has to pay the price and can potentially gain in profit, if they pass on this opportunity there is no cost nor profit.

Company’s IT team is not so experienced and the configuration of software is complicated. Because of this, if the company decides to buy the software, there is only 15% chance, that configuration will be successful and company will gain $60.000 thanks to it.

If IT department fails to implement the software, the company can hire external consultants to help them. Cost of such service is additional $15.000. Unfortunately, the software is a new product so there is only one company which offers installation of it, and it is a new company with no reputation – therefore there is a 7% chance that even consultants will fail to configure software correctly. If consultants do their work, our investment fund will gain $60.000, if not, there will be no additional gains/losses.

Based on provided information please prepare decision tree using SilverDecisions (<http://www.silverdecisions.pl/>) and based on it suggest whether investment fund XYZ should buy the software? If yes, should it be prepared to hire consultants? In other words, provide optimal decision path. What is company’s expected profit in terminal node which has the highest probability of occurring? Company wants to maximize its profit.

**Please write your answer in R script as a comment and additionally send .pdf file with exported tree (include your student ID in one of nodes’ labels as a proof of work).**

**Ex. 2 (13 p)**

Work with **Carseats** data from **ISLR** library. In this exercise you will be asked to provide a model with *Sales* as dependent variable.

1. (1.p) Prepare a plot presenting the distribution of Sales’ variable.
2. (2p) Divide data into training (80%) and validation (20%) sets (with seed equal to student ID).
3. (2p) Prepare a linear regression with *Sales* as dependent variable and *Age, Price, Income* as explanatory variables on the training set. Check which variables are statistically significant (write answer as comment).
4. (3p) Prepare a linear regression with *Sales* as dependent variable and *Age (polynomial of 6th degree), Price, Income* as explanatory variables on the training set.
5. (2p) Prepare a random forest model (based on 10 trees) with *Sales* as dependent variable *Age, Price, Income* as explanatory variables on the training set.
6. (3p) Compare quality of the models on validation set: calculate MSE for three models mentioned in previous points. Which model is the best (write answer as comment)?

**Ex. 3 (25p)**

Use ‘**college\_test.csv’** dataset with statistics for US Colleges. Your dependent variable is ***Private****,* which takes value 1 for private college and 0 for public college. Values in the dataset are separated by comma and dot is a decimal point. First column contains row names.

1. (1p) Check if there are any missing values in the dataset.
2. (2p) Check the percentage of observations where college is private (write answer as comment).
3. (2p) Prepare a plot presenting the distribution of *Private* variable.
4. (2p) What is the mean value of variable *Expend* for *Private* equal to 0 (write answer as comment)?
5. (1p) Divide data into train (75% of observations) and validation (25% of observations) sets (with seed equal to your student ID).
6. (1p) Prepare a decision tree with rpart package on training set (dependent variable is *Private* and your goal is to predict if school was private; take cp=0.00001, minsplit=2, rest default).
7. (3p) Find optimal value of cp parameter using cross validation (write optimal value in comment).
8. (2p) Prune your tree for optimal value of cp.
9. (3p) Write one decision rule for unpruned tree (from top of three to one of terminal nodes, write as a comment).
10. (5p) Check model quality (pruned) on the validation set: prepare confusion matrix, calculate model accuracy and true positive rate and false positive rate.
11. (1p) Draw a ROC curve for pruned tree.
12. (2p) Check the AUC and interpret it (write as a comment).

**Your submission should contain:**

* **pdf with the decision tree from SilverDecisions (with your student number in one of nodes’ labels)**
* **executable R script with codes used and answers to questions as comments where specified.**

**Please send your submission via e-mail on:** [**nosarzewski.aleks@gmail.com**](mailto:nosarzewski.aleks@gmail.com)**. Late submissions will not be accepted.**

**Good luck!**