

Questions Q&A 4 Decision Trees – Tuesday Nov 23rd

Coding Questions – Gijs Breakout Room

- [Lucrezia];

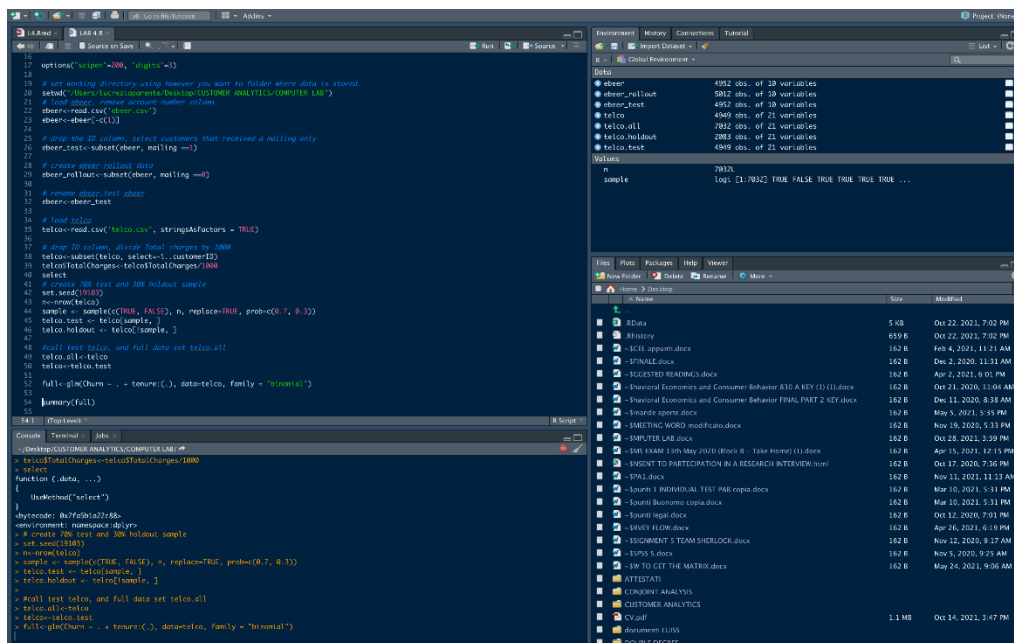
I am having some problems running the code we did in class during the Lab.

If I run this:

```
full<-glm(Churn ~ . + tenure:(.), data=telco, family = "binomial")
```

I don't get any result, just the stop label in the console. I've waited for a long, but still it doesn't work.

Can you help me?



```
16 options("out" = "Rplots.pdf")
17
18 # Set working directory using wherever you want to folder where data is stored
19 setwd("D:/r/courses/decisiontrees/DECISION TREES/COMPUTER LAB")
20 # Load data, remove dataset number column
21 telco<-read.csv("telco.csv")
22 # Create test set
23 telco.test<-telco[1:1000,]
24
25 # Drop the 20 columns, select columns that received a mailing only
26 telco.test<-subset(telco.test, mailing==1)
27
28 # Create other rollout data
29 telco.rollout<-subset(telco.test, mailing==0)
30
31 # Remove other test data
32 telco.test<-telco.test[1:1000,]
33
34 # Load data
35 telco<-read.csv("telco.csv", stringsAsFactors = TRUE)
36
37 # Drop 20 columns, divide total charges by 1000
38 telco<-subset(telco, select=1:customerID)
39 telco$totalCharges<-telco$totalCharges/1000
40
41 # Create test and train foldout sample
42 set.seed(12345)
43 n<-nrow(telco)
44 sample<-sample(c(TRUE, FALSE), n, replace=TRUE, prob=c(0.7, 0.3))
45 telco.test<-telco[sample,]
46 telco.rollout<-telco[!sample,]
47
48 # Split test telco, and full data set telco.all
49 telco.all<-telco
50 telco<-telco.test
51
52 full<-glm(Churn ~ . + tenure:(.), data=telco, family = "binomial")
53
54 January(Full)
55
```

Questions Week 3 – Logistic Regression

- [Mieke];

(1) In Q3 of the practice quiz, why do we use Multiple R-squared and not Adjusted R-squared?

(2) In Q4 of the practice quiz, 'How much more or less likely are the customers in the high tenure group to churn relative to the low tenure group?', I do not understand why this gives the correct answer to the question: $\text{round}((\exp(\text{coef}(\text{model_4})["\text{tenure_group3}"])-1)*100,0)$.

This is not relative to the low tenure group or is it?

Questions Week 4 – Decision Trees

- [Mieke];

(1) In Q1 of the practice quiz, why don't we first have to mutate Churn to 0,1 values?

(2) In Q2 of the practice quiz, isn't it `InternetService = as.factor(2)` instead of `InternetService = as.factor(1)`?

- [Noa];

(1) When doing a random forest, should we always set the seed or not?

(2) When choosing the tree with the min. OOS deviance, we look at the plot with OOS deviance and tree size and see which tree size is the best option. What should we do when this graph is not clearly readable? Is there another way to get the value for the tree size?

(3) Regarding Practice Quiz - Question 2: In the script for the quiz we use `newdata1 = data.frame(gender = as.factor(2) ...` However, female is coded as the first level in the data

```
> levels(telco$gender)
[1] "Female" "Male"
```

So I was wondering, shouldn't we then use `gender = as.factor(1)` or is there another reason as to why we use `gender = as.factor(2)`. Same question holds for the `PaymentMethod`-variable.

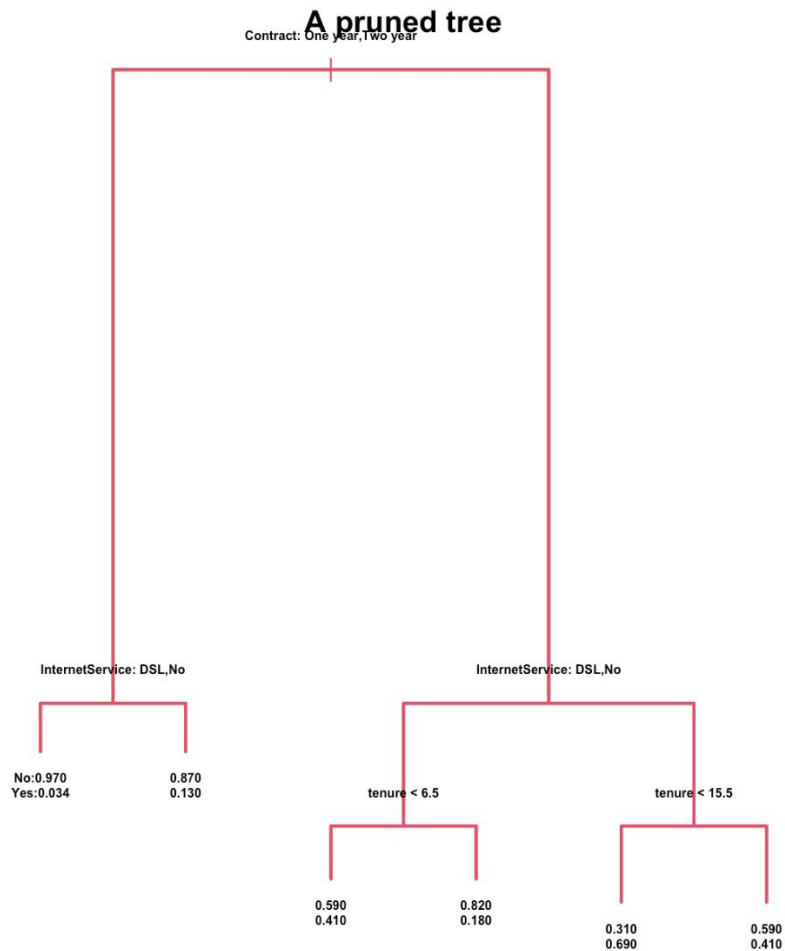
(4) Also in question 2, we should focus on a one year contract. The code is:

`Contract=as.factor(levels(telco$Contract))[2]`. Why do we use a different formula here? It seems that we should do the same thing as before, right? So why not just `contract = as.factor(2)`?

- [Kevin];

(1) Regarding practice quiz Q5, could you explain why the answer is 0.034 instead of 0.130?

When looking at the tree, I understand that the first decision in the tree is to go left, but I do not understand why at the second node (Internet Service) you also have to choose the left side (As in the text 'no' is on the right). I have also included a png of the tree;



(2) Regarding Q9 of the assignment, the instruction states: "Use 10 fold cross-validation to prune the tree". Does K-fold cross validation mean we always have to start from the most complex tree or does it in this question mean we have to start from the tree estimated in Q7?

(3) Also for Q9, I do not get text shown in the decision tree. Do you know why this is the case?
This is the code I used:

```

par(mfrow=c(1,2), oma = c(0, 0, 2, 0))
tree_cut<-prune.tree(tree_complex, best=4)
plot(tree_cut, col=10, lwd=2)
text(tree_cut, cex=1, label="yprob", font=2, digits = 2, pretty = 0)
title(main="A pruned tree")

```

A pruned tree

