INFT216 - Data Science 193

Assignment 1: Predicting the success of Bank Telemarketing using Decision Trees

Assessment Value: 10%

Due Date: 5pm Sunday 27th October 2019 via iLearn

This assignment uses actual data from a Portuguese bank which uses a data driven approach to predict the success of telemarketing calls for selling long-term bank deposits.

You can read much more about this data (released to the UCI Machine Learning community) here: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing# . There are also big data/data science papers using this data, for example this one in 2014: http://www.sciencedirect.com/science/article/pii/S016792361400061X

I have loaded this data into an SQL database (to make sure you get some practice using SQL!). The data is located at BRUCEDBA.BankMarketing, and you should access it directly from R following the same procedure we used in workshops.

Here is the brief description of the field names in the SQL database (note some are subtly different from the UCI names at the link above).

Input variables:

# bank client data:

1 - age (numeric)

2 - job : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","selfemployed","services","student","technician","unemployed","unknown")

3 - marital : marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)

4 - education (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown")

5 - defaultcredit: has credit in default? (categorical: "no","yes","unknown")

6 - housing: has housing loan? (categorical: "no","yes","unknown")

7 - loan: has personal loan? (categorical: "no","yes","unknown")

# related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: "cellular","telephone")

9 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

10 - day_of_week: last contact day of the week (categorical: "mon","tue","wed","thu","fri")

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

# other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

```
13 - pdays: number of days that passed by after the client was last contacted from a previous
   campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical:
"failure","nonexistent","success")

   # social and economic context attributes

16 - emp_var_rate: employment variation rate - quarterly indicator (numeric)

17 - cons_price_idx: consumer price index - monthly indicator (numeric)

18 - cons_conf_idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr_employed: number of employees - quarterly indicator (numeric)


  Output variable (desired target):

21 - y - has the client subscribed a term deposit? (binary: "yes","no")
```

Your goal is to create a decision tree which can predict class membership of the “y” variable. Clearly, the better the tree prediction, the happier management will be!, and to create the best model you can to predict the probability a customer will take up the offer. You will need to be able to assess the models ability to predict on unseen data. The marketing team will then use this model, and thus your probability estimate, to decide who to mail information out to. Clearly, the better the prediction, the less the cost of running the campaign.

Deliverables:

Your final deliverables will be 2 PDF files, both produced by the same .Rmd script (with different code chunk options). You must submit:

PDF1 - all code and results shown (like you would share with a colleague on the Data Science team)

PDF2 - only show those things necessary to help support management decision making (this is the one you send to management!)

These will both be submitted online through iLearn.

Helpful hints:

- Think about what you are trying to build/show ○ Focus on good document structure and layout (revisit week 2 on repeatable research)
    - Hint: Think about the headings in the document you produce
    - You need to make your own R functions (using the function command – see our examples in previous weeks on iLearn):
        - Your own function for creating the decision tree (you will call the R function to create the tree from inside your new function), and,
        - Your own function for producing a prediction from your tree.

        - (These will come in handy later in the course!) – don’t just call the existing tree functions directly from the script!
    - Read the data in from the SQL database using SQL
    - You will need to convert the attributes to their correct datatypes and factors etc

- You may use any [R] package to build the tree

  • You may build more than 1 tree and compare them ○ You may also prune trees if you wish

  • If you build more than 1 tree, you will need to compare trees and explain the difference. Also, you will need to clearly recommend which tree is best and why.

  • You should use the GLM package to build the logistic model.

- Focus on letting the visualizations do the talking. Only include explanatory text where it is really necessary… although you should remember that management do not really understand data science, so you will need to find a tradeoff between understandability and verbosity. Verbose assignments will be penalized.
    - You will need to use visualizations
    - You will need to be able to show how good your tree(s) is/are at classifying ○ You will need to show some example predictions from new data that you create yourself

  Note:

As is the case with all assignments I set, if you do the minimum (correctly), then you will receive half marks. Additional marks are awarded for those assignments where you have clearly put in additional thought, whether it be in visualization, modelling, succinctness, or coding elegance.