INFT216 - Data Science 191

Assignment 1: Understanding Customer Churn at BondTelco

Assessment Value: 10%

Due Date: End of Week 7, Sunday 3rd March 5pm (submit via iLearn)

Congratulations, you are a newly employed data scientist at BondTelco. BondTelco is a retail provider of contract mobile phone services. Although quite a large company, BondTelco have not yet dabbled in Data Science, indeed, you are their first employee in this role.

Currently, management at BondTelco are concerned about the high rate of churn among their customers. To try and address this concern, the sales staff have been encouraged to ring up customers whose contracts are coming due, and offer them incentives to stay with the company. Unfortunately for management, this is an expensive process, as it involves offering incentives to all customers, whether they are likely to leave or not. What is really required is a way to predict whether a given customer is likely to leave or stay with the company. In this way, only customers who are likely to leave will be offered the incentives, thereby reducing costs.

The first question you have been asked to address is: ‘Is there a way to determine in advance which customers are likely to leave when their contracts are up?’

The IT team provide you with access to the company database, which contains a table (BRUCEDBA.BondTelco_Customers) containing data on 20,000 previous customers, including whether they left or stayed with the company at the end of their contract period. The data in this table are:

COLLEGE	: Is the customer college educated?
INCOME	: Annual income
OVERAGE	: Average overcharges per month
LEFTOVER	: Average % leftover minutes per month
HOUSE	: Value of dwelling (from census tract)
HANDSET_PRICE	: Cost of phone
OVER_15MINS_CALLS_PER_MONTH	: Average number of long (>15 mins) calls per month
AVERAGE_CALL_DURATION	: Average call duration
REPORTED_SATISFACTION	: Reported level of satisfaction
REPORTED_USAGE_LEVEL	: Self-reported usage level
CONSIDERING_CHANGE_OF_PLAN	: Was customer considering changing his/her plan?
LEAVE	: Whether customer left or stayed

Your goal is to create a decision tree which can predict class membership of the LEAVE variable. The sales team will then use your tree’s rules to determine whether a customer is likely to stay with or leave the company. Clearly, the better the tree prediction, the more retention costs can be reduced.

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.

Note: Everything in R takes much longer than you think it will! My advice is to start promptly! I cannot give you an extension because you didn’t start early enough… its not fair to other students!

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 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.

- 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.