Assignment for Data Analytics

\*\*Character vars as factors

The objective of this assignment is to use some of the techniques we have covered in class. This year each person will focus on one technique. Two datasets will be supplied one with a binary outcome and one with a continuous outcome. You are expected to use both.

**Preliminary Analysis.**

Everybody is expected to explore both datasets and summarise the results. For the sake of brevity you are required to look at 3 independent variables in each dataset vs target variable. These 3 variables should be a mixture of continuous and categorical variables. Obviously you should review all variables.

The next part of the assignment consists of choosing a topic from the list below. Each topic will involve applying the basic technique as covered in lectures. **All models should be evaluated appropriately.** For each choice you are also required to carry out some additional research.

**Choice 1 Growing trees**

This topic consists of the following tasks

1. Grow and evaluate a simple tree using the CART technology for each of the datasets
2. Derive and implement a different splitting criteria (one for use with binary outcome variables and one for continuous variables). See vignette in R for how to write new splitting functions. Compare the results to the trees grown using the CART methodology. You should justify your choice of splitting criteria.
3. Review and implement one other approach of growing trees – available in R.

**Assignment of marks**

Basic Summary Statistics (Graphs & whatever) 10%

Grow and evaluate a tree on each of the datasets 40%

Different Splitting criteria 25%

Other approach (part 3) 25%

Choice 2 **Random Forests**

This is a classical ensemble method which will be covered in class.

Apply the basic random forest method to the two data sets. Include a detailed description of setting any parameters and the various outputs produced.

Random Forests weight each of the trees equally. Implement the method allowing for unequal weighting of trees.

Explore other types of random forests open to suggestions here e.g. generalised random forests

**Assignment of marks**

Basic Summary Statistics 10%

Grow and evaluate a random forest on each of the datasets 40%

(This includes tuning any parameters)

Unequal weighting of trees in forest 25%

Other types of random forests 25%

**Choice 3 Boosting algorithms**

Carry an extensive exploration of the two boosting methods gradient boosting and Xgboost. This involves exploring the various parameters involved and how they affect the solution and also a description of the various outputs produced.

**Assignment of marks**

Basic Summary Statistics 10%

Gradient boosting model 40%

Xgboost 50%

Date of submission 12.00 on Monday December 3rd

Please hand a hard copy to School reception in the O’Reilly Building and submit a copy to Turnitin via Blackboard.

Figure 6: Proportion of Churn:Remain for each Tenure Group

Figure 5: Histograms of Count & Proportion of each Tenure Group

Figure 4: Summary of Tenure Group