# BUS212f2: Analyzing Big Data II Spring 2017

ANALYSIS 2: ~/Desktop/Brandes/Analysing Big Data 2/Analysing Big Data 2/data/Betting.csv2

### Introduction

In the case “A Game of Two Halves” we find a presentation of two common classification techniques applied to a large sample of English Premier League football results. The goal of the two approaches is the same: to see if bettors can use data from the first half of a soccer (football) match to predict the ultimate outcome of the match. The two techniques are multinomial logistic regression and a decision-tree method called chi-square automatic interaction detection (CHAID).

In this case, the authors have created predictive models using the two techniques, and our task is to understand and apply these techniques to a fresh set of data to evaluate their performance. In the course of doing so, you will deepen your understanding of the techniques as well as some important fundamental concepts of data analytics.

### Software and Data:

On LATTE you will find two Excel data files. The first is called “Betting.csv”; this contains the relevant data from the 1,520 matches referenced in the case. You should use R to do this analysis, as we began to do in class. The second CSV (“BettingTest”) file contains 20 “new” matches to analyze later in the assignment.

You should treat the first file as your Training sample, and the second as the Test sample.

### Questions:

You will find the questions for this case on page 4 of the case itself. There are 10 in all, and I’m substituting the following questions for the corresponding ones in the case. For **Questions 1­­--4, use your R output in place of Exhibit 7**:

|  |  |
| --- | --- |
| Replacement for Question | Replacement wording |
| 2 | According to your model, how do red cards conceded by the Home and Away team influence the outcome of a match. Your response should not be numeric, but qualitative. Based on your understanding of the sport, speculate about possible reasons for these findings. |
| 3 | Under what circumstances is it informative to include points scored by a team in the prior season to predict the outcome of a match? |
| 4 | Using the multinomial logit model, compute the probability of a home team win for a match with these attributes: |
|  | |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | HTGD | FGS | RED\_H | RED\_A | POINTS\_H | POINTS\_A | TOT\_H\_P | TOT\_A\_P | | 3 | 1 | 0 | 1 | 18 | 15 | 39 | 32 | |
| 7 | Using the Decision Tree from our R Script, write out several practical rules ***for*** betting. Be selective; if there is a tree branch that essentially says “don’t bet at all”, then omit that rule. |
| 8 -- 10 | For these, use the R output instead of the Exhibits mentioned in the case. |

During class, we will develop the R code to replicate some of the results shown in the case. Prior to responding to any of the questions, be sure to get the scripts running on your own computer, using the initial case data. Remember that the code I distributed has the directory path for my computer; you’ll need to modify that for your won.

Additionally, you’ll need to create and “knit” an R Markdown file that includes your r code chunks, output and your commentary and responses to the questions.

**NOTE: Knit the file as a Word document, and then re-save the Word doc as a pdf file.**

### Deliverables

As usual, you may work alone or in a small team. If you work in a team, please upload just *one* copy of the pdf file before the due date. Your submission for this Analysis will consist of one pdf uploaded to LATTE.