BIA 6301 APPLIED DATA MINING

HOMEWORK ASSIGNMENT #5

General Instruction: The homework assignment is due one week after it is assigned (@ 5:45 PM on Tuesday). Assignment turned in after the due date and time will lose 2 points for every day late. No assignment will be accepted one week after it is assigned. Here’s a breakdown of the point distribution.

|  |  |  |
| --- | --- | --- |
| **Task** | **Points Possible** | **Submission Instruction** |
| GitHub repository | 1 point | Submit the hyperlink to the GitHub repository & the html viewer link of the knitted html file onto Blackboard for grading. |
| Link of html viewer of knitted markdown file | 1 point |
| Markdown file with written commentary. | 3 grading parameters x 6 points = 18 points  The markdown file will be graded using the Written Assignment Grading Rubric. |
| ***Total*** | ***20 points*** |  |

GitHub & HTML Viewer

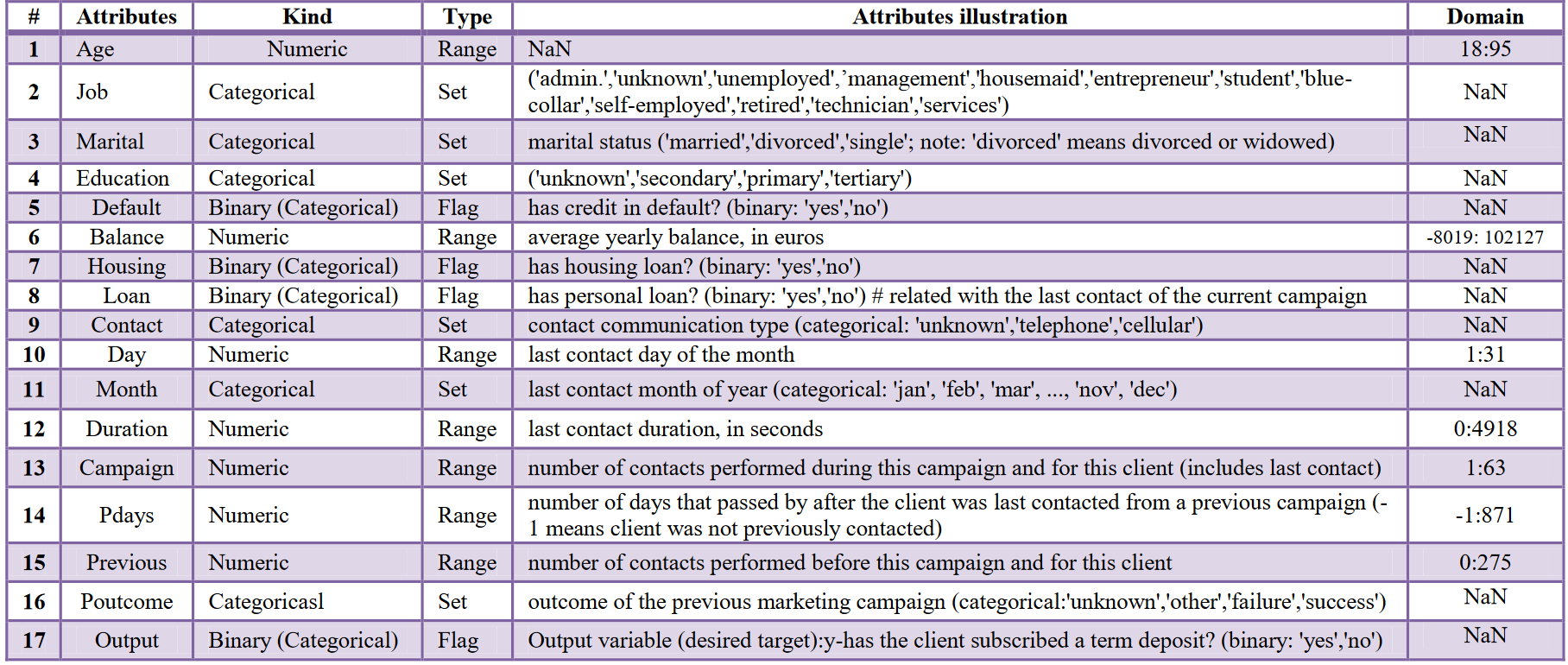
You will create another detailed markdown file. Please submit two links for grading this week:

1. GitHub repository for this week’s assignment. I want to see the data, Rmd file, and html output of the knitted markdown file.
2. Hyperlink to view your knitted markdown file in html format. Append the following link in front of the hyperlink to the knitted html file posted on GitHub: **http://htmlpreview.github.io/?**

For example, here’s the html view of the markdown file from last week’s class: <http://htmlpreview.github.io/?https://github.com/xhp4y8/Applied_Data_Mining/blob/master/PCA.html>

**Case Study: Improving Term Deposit Client Base**

You have been hired by Hometown Bank to help the company understands and improves its status in the term deposit market. The bank has given you two different versions of a data set: a large version with 40,000 rows (**bank-full.csv**) and a sampled, smaller version with 4,000 rows (**bank.csv**). The bank would prefer you to use the full data set for your analysis but understand if technology limits mean that you can only use the smaller one. The variable the bank seems most interested in understanding is “y::category”, which refers to whether the client has subscribed a term deposit. It is a factor with two levels: “yes” or “no.” The following table shows the variables in the data set:



Source: Elsalamony, Hany (2014). Bank direct marketing analysis of data mining techniques. *International Journal of Computer Applications* 85(7): 12-22.

Just as you are getting ready to dive into the data, you meet a colleague who has done consulting for Hometown Bank in the past. She gives you the following advice:

Make sure they understand clearly the business question you have investigated and the way you approached the answer. The bank president is a very nervous person who will insist that you include multiple approaches in your analysis so he knows that you have covered all the bases. He won’t listen to you if you don’t have **at least three different models to back up your suggestions**. Furthermore, the vice president of the bank is very skeptical of data science and will be looking for ways to discredit your work. Make sure you clearly explain how the data are handled, the models tested, and the recommendations arrived at or he will not believe you. Beware of jargon – it will definitely put the vice president off.

Undertake this analysis and provide the results to your client.

**General Advice on How to Proceed**

I will grade your knitted markdown file with commentaries using the Written Assignment Rubric as usual. You should address all three grading parameters in your markdown file: 1) What is the problem?; 2) What have you done to address the problem?; and 3) What are your recommendations?

The grading distribution for each grading parameter is as follows:

Exceeds Expectations = 6 points

Meets Expectations = 4-5 points

Below Expectations = 2-3 points

Fail = 1 or 0 points

You have worked hard learning many different models in ADM in the past six weeks. Use this assignment as an opportunity to showcase what you have learned. Good luck!