BIA 6301 APPLIED DATA MINING

HOMEWORK ASSIGNMENT #1

Due 5:45 pm on March 28

General Instruction: 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 is a breakdown of the point distribution for this assignment.

|  |  |  |
| --- | --- | --- |
| **Task** | **Points Possible** | **Preferred File Name** |
| Part A: 2 questions  Part B: 4 questions | 2 x 1/each = 2 points  4 x 4/each = 16 points | LastName\_text.doc  (or .pdf) |
| R markdown documentation for Part B | 2 points | LastName\_code.html (or .doc or .pdf) |
| ***Total*** | ***20 points*** |  |

To help facilitate the grading process, please use the file naming convention listed in the table above. Please upload the required files onto Blackboard for grading.

The markdown files will not be graded, but they will be checked if necessary to verify your findings and recommendations. Point deductions may occur if there are major discrepancies between your written answers and the R markdown files.

**Personal Loans at Universal Bank**

**Part A.**

You were recently hired as a data scientist by Universal Bank. The bank’s Vice President is interested in building a model to help predict when users might respond to a campaign to take out a personal loan. She supplies you with a dataset containing information of 5,000 customers and attaches a description about the dataset (see page 3 of this assignment).

The VP tells you that another analyst has created a logistic regression model and hands you an R code file (please download the Template\_HW\_1.Rmd file). She has some questions about the code file:

1. What does this code do? What does the output mean? How can you use this to know what options you might explore for expanding Personal Loans?
2. A new customer comes into the bank to get information about refinancing her mortgage. She is single and 38 years old with 17 years of experience. She has an MSBIA, makes $150,000 per year. She spends $200 a month on her credit card with Universal Bank, usually using online banking. She doesn’t have a mortgage, a CD or a securities account and is the only member of her family that banks with Universal. Would this model predict she would be interested in a personal loan?

**Part B.**

Now that the Vice President understands what the analyst did, she is looking for you to create other models with which she can compare them. Run a KNN model, Naïve Bayes model, and a decision tree model. Use your results to answer the following questions:

1. Explain how each model works and how you set up the problem. What decisions did you make? Why?
2. What does the output mean? How can you use this information to know what options Universal Bank may want to explore for expanding personal loans?
3. How do these models agree with each other and with the prediction of the logistic regression model for our new customer example?
4. Which model has the highest predictive accuracy? Which model would you recommend using? Why?

Please provide the answers to the VP’s questions in clear, concise paragraphs. For Part A, there should be no more than one paragraph. Part B should take no more than one paragraph per question or 4 paragraphs total. Since you are writing to someone who has a strong quantitative background, you can use technical terminology—just be sure that you are using it correctly!

Please also knit your markdown file and submit it with the text. The markdown file will not be graded, but it will be checked if necessary to verify your findings and recommendations. **(2 points)**

**Description about UniversalBank.csv file**

The file **UniversalBank.csv** contains data on 5,000 customers. The data include customer demographic information (age, income, etc.), the customer’s relationship with the bank (mortgage, securities accounts, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5,000 customers, only 480 (9.6%) accepted the personal loan that was offered to them in the earlier campaign.

Here is a description of each variable in the Universal Bank dataset:

ID: Customer ID

Age: Customer’s age in years

Experience: Number of years of professional work experience

Income: Annual income in thousands of dollars ($000)

Zip code: Zip code of home address

Family: Customer’s family size

CC Avg: Average spending on credit cards per month in thousands of dollars ($000)

Education: Education level where 1 = Undergraduate; 2 = Graduate; and 3=Advanced/Professional

Mortgage: Value of house mortgage if any; in thousands of dollar ($000)

Personal.Loan: Did the customer accept a personal loan offered in the bank’s last campaign? 1=Yes; 0 = No.

Securities.Account: Does the customer have a securities account with the bank? 1 = Yes; 0 = No.

CD.Account: Does the customer have a certificate of deposit (CD) account with the bank? 1 = Yes; 0 = No.

Online: Does the customer use Internet banking facilities? 1 = Yes; 0 = No.

Credit.Card: Does the customer use a credit card issued by Universal Bank? 1 = Yes; 0 = No.