BIA 6301 APPLIED DATA MINING

HOMEWORK ASSIGNMENT #3

General Instruction: The homework assignment is due on the assigned date at 5:45 PM. 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.

|  |  |  |
| --- | --- | --- |
| **Task** | **Points Possible** | **Preferred File Name** |
| Part A: Association rules  Part B: Model evaluation  Part C: Final Project Proposal | 3 X 4/each = 12 points  3x3/each = 9 points  2 points | LastName\_text.docx |
| R markdown documentation for Parts A & B | 2 points | LastName.html or LastName.docx or LastName.pdf |
| ***Total*** | ***25 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 html outputs of your markdown files will not be graded, but they will be checked if necessary to verify your findings and recommendations. Raw markdown (.Rmd) files will not be accepted. Point deductions may occur if there are major discrepancies between your written answers and memorandums and the knitted markdown files.

Please upload your Word documents individually onto Blackboard. Please put your html files in a zipped folder and upload it. Blackboard’s upload feature does not accept html files.

**Part A: More unsupervised learning**

You have made some suggestions to the CEO about pricing or offering options. She is interested in your suggestions but would like more analysis to be convinced.

Try using association rules with the crash data. In no more than 3 or 4 paragraphs, answer the following questions for the CEO about the results of this additional analysis:

1. How is the analysis different than what you presented before? What changes, if any, did you have to make to the data?
2. Did this analysis give you any insights that you didn’t have before? If so, what are they? If not, why do you think that might be?
3. Does this new work support your previous recommendations? If so, how? If not, what recommendations would you make now? How confident are you in the results?

REMEMBER: this is a **general language response**. Use of jargon is discouraged and use of *unexplained jargon* will result in loss of points.

NOTE: The **crash.csv** data set was taken from Fatality Analysis Reporting System (FARS) of the National Highway Traffic Safety Administration of the U.S. Department of Transportation. Here is the link to FARS: <https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars>. If you feel that you would benefit from having additional data for this exercise, you are welcome but **not required** to do so.

**Part B: Model evaluation**

Revisit the work you did for Homework 1. Building on the code you already have, consider several (at least 3) changes you would make to your models. In no more than 3 or 4 paragraphs, answer the following questions.

1. Consider/perform a variety of evaluation measures for the work you have already done. Is accuracy the best metric for performance for your question? If so, why? If not, why not? What measure would be better? How does your recommendation from Homework 1 change?
2. What changes did you make to the original models? Why? What do you expect to see from these new specifications? What are the results?
3. Compare the entire set of models using your preferred criteria. Which performed best? Which would you recommend? Why?

The audience here is other data scientists. Technical language is fine as long as it is used correctly.

**Part C: Final Project Proposal**

**NOTE:** While this is not due until class time in week 7, you are welcome to submit just this one paragraph at any time if you would like feedback earlier. This means you MAY submit it early, not that you MUST.

Please write one paragraph describing your final project proposal. Please include these items in your project proposal:

* Business question(s) you want to answer or solve
* Data source(s) you plan to use
* Supervised and/or unsupervised learning methods you plan to use to mine the data source(s). A minimum of two methods is required.