Data Mining
IT 270
Final

Prof. Alexander Pelaez
apelaez@hofstra.edu

$1^{st} May, 2023$

Instructions

Please make sure if you use R you copy and paste it into Word using Courier Font (makes it easier to Read). For each of the problems that are looking for a response (not just a calculation), be sure to explain and interpret the results. If you aren't sure. . . ASK PLEASE.

Please start each question on a new page and clearly label that start of each problem (Maybe slightly larger font, bold face , underline. . .) anything that will help me find the problem you are working on.

As a final exam you are to work on your problems individually. However, you may discuss techniques and approaches. You may not copy code or answers and copying other students code and answers could result in major penalty or even failure on the exam or the class

1 People Analytics (GoT...Analytics) (30 Pts)

Files Needed: Same as midterm

The CEO is concerned regarding recent events at the company, and wants to initiate a new program. It is concerned with gender discrimination and equitability across the board. They are seeking to also increase overall job satisfaction.

Using the employee datasets (from the midterm), provide a determination if there is gender discrimination in the workplace, and provide an assessment of factors that should be looked at for job satisfaction and how these factors can be used. This should be a very comprehensive review.

A few notes: You must come up with appropriate methodologies and applications of those methodologies. Your final answer should be in the form of text with supporting facts (maybe from your analysis in R). Expect this to be read as a report by the CEO. At the end of your answer you can provide all necessary R code and outputs.

Any points you raise must be justified. If you try something let's say you wanted to determine if employees that have dogs instead of cats are important. You should list it and try and say it didn't work...make sure it was relevant in the first place. You may write even something as small as a line or two on that. If something is important...write about it.

You may ask me any question you like on this and I will answer it as best as I can. If you ask, how many methods do I need. . . answer from CEO (I don't know you're the expert...that's what I pay you for). You may use techniques from BAN203 or BAN250, but they cannot be the core of the paper, they can only be used as support or justification.

2 Handwriting Recognition: ENDGAME (30 Pts)

Your colleague was called into the bosses office and was reprimanded. Your colleague argued that your prior analysis is wrong and you didn't explain anything about the model.

(a) Why was your colleague wrong in the first place?

(b) Your colleague reran the code and claims that a one layer ,40 hidden node model actually works best, since there are ten numbers and each of the nodes gives a 25% probability of hitting a number from 0-10, and thus each hidden node can be explained clearly. Explain why this is a good rationale or not

(c) Re run your neural network, however, this time you have a training set of 60,000 and the test set of 10,000 (this may take some time to run).

- Provide a table and graph of the error rates in your training set and test based on the number of nodes chosen in your layers. Since there may be quite a number of combinations, be effective in how you approach this.
- Provide your final answer, with the error rate and time it takes to complete the analysis.

(d) Consider all the techniques we learned in class

- Can any other technique be used to perform the classification. If so list the techniques with a one-line rationale as to why
- Run the techniques in part 1, and provide the final answer for that technique along with the training error rate and test rate (be sure to indicate any other trials you had).

(e) Consider everything above, write a summarized conclusion of everything above you would hand to a Chief Analytics Officer.

(f) After this, what should happen to your colleague?

3 Performance Optimization (Return of the FIFA) (30 Pts)

The FIFA Director of Analytics isn't completely convinced with your prior analysis. The belief is that your analysis isn't useful. Use the original fifa.csv dataset.

(a) Explain how your prior FIFA analysis was useful and how it can be used.

Ultimately there needs to be a unifying factor across all techniques used. If the primary question is how to build a team then your job will be to develop a methodology that will assist in this, therefore:

(a) How would you group players ? Be sure to establish what fields are necessary prior to grouping and indicate how you grouped them. Then describe the groups.

(b) Using the groups can you assess the clubs. Explain if this is useful.

(c) How can your initial analysis (from midter) be integrated with the previous two bullet points?

(d) Can you find any other useful and interesting analytics (from our techniques) that would be helpful to the Director of Analytics.

4 Question 4 - Theoretical Questions (10 Pts)

Provide no more than a paragraph for each, be concise.

(a) Michele Piccinno a researcher in Artificial Neural Networks stated (2016), “Neural Networks are the second best way to solve a problem. The best way is to understand the problem”. Provide an opinion based on your knowledge of whether you agree or disagree with this statement?

(b) Explain how accuracy is measured across the techniques and why is it appropriate forthose techniques?

(c) What are some of the challenges with data when starting a project and what would your initial steps be?

(d) Explain how a decision tree can be used to predict a continuous variable. How are the results interpreted?

(e) Provide an example of an ethical problem with data mining, and what an analyst should consider and how they should mitigate the problem, if possible