week15_coursework

February 13, 2025

1 PHYS465: Coursework Exercise 4: Part 2

1.0.1 Deadline Monday 17th Feb @ 4pm

1.0.2 This coursework assesses the learning outcomes from Week 15.

The final coursework exercise for PHYS465 is made up of 2 parts. This is part 2. Your submission should include both exercises.

1.1 Instructions

- Submit your work via Moodle.
- You must submit a fully compiled `.ipynb` file which includes all codes required to replicate your results **and** a .pdf version.
  - **Dont forget to run every cell before submitting**
  - You must also respond to the mandatory GenAI self-assessment questionaire.
- Your submission must include text (in markup format) that describes what each cell does and summarises the conclusions
- The estimated workload for this is 4-6 hours.

1.1.1 Tips

- The last question of this exercise asks you identify a key result. **To do this you do not have to have completed all exercises**. This assessment is designed to test your reflections on the problem undertaken.
- Don't worry too much about how your code looks - while some marks will be given for sensible coding, the focus of this assessmnt is your approach used in solving the problem, your reasoning, explanation and answer.
- As data visualisation is a key outcome, marks will be given for well presented plots
- Explain all your reasoning for each step. A *significant fraction* of the marks are given for explanations and discussion, as they evidence understanding of the analysis.

- Include all relevant lines of code including import statements and read statements. As part of the assessment your code will be run offline.

1.1.2 WARNING

- This submission must be your own work. Please note the university's policy on plagiarism.
- While it is acceptable (and indeed encouraged) to share ideas, you must ensure that you do not use other people's code or text, and that the reflections are your own.

- It is acceptable to use GenAI tools for guidance on how to approach this exercise, but you must ensure that all code is written by you.
  - Should you use GenAI in this work, then answer yes to the GenAI self-assessment. You will not be penalised for this. ***

1.2 The Problem

Dataset 1:

For this exercise, we will use data from the Titanic to predict the likelihood of survival.

Two datasets are available: * Training data: `https://raw.githubusercontent.com/MatSmithAstro/phys465_re`
* Test data: `https://raw.githubusercontent.com/MatSmithAstro/phys465_resources/main/coursework/da`

For each of the two datasets, the following variables are available: * `survival`: Survival (0 = No, 1 = Yes) * `pclass` : Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd) * `sex` : Sex (or Gender) * `Age` : Age in years * `sibsp` : Number of siblings / spouses aboard the Titanic
* `parch` : Number of parents / children aboard the Titanic
* `ticket` : Ticket number * `fare` : Passenger fare
* `cabin` : Cabin number
* `embarked`: Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton) ***

1.3 Part 1: Dataset Analysis

1. Read the dataset into a pandas dataframe and identify the features for training
   [2 mark]
2. Make a multi-panel plot to show the relationship between survival and the key parameters of interest for the training dataset
   [5 marks]
3. Make a multi-panel plot to show how representative the training dataset is of the test dataset.
   [5 marks]
4. For the parameters that you have decided are importance predictors of survival, recast the parameters to be between [0,1], or so that they are suitable for a boosted decision tree model

- *Hint*: do not forget to replce blank entries
  [3 marks]

5. Split the training data in two, saving 25% of events for model testing
   [3 marks]
6. Create a booted random forest model to predict `survival` given your parameters of interest
   - *NB*: the tree should be optimised using the AUC metric.
     [5 marks]
7. Determine the importance of all feature's considered, and calculate the key metrics (accuacy, precision and recall) for the final model
   [5 marks]
8. Visualise final model
   [4 marks]
9. Make predictions for all members of the test dataset. Visualise the results.
   [4 marks]
10. Predict the outcomes for two fictious passengers : Jack, a 20 year old, 3rd class passenger, and Rose, a 17 year old first class passenger.

[1 mark]

Maximum possible 35

1.4 Part 2: Extension and Summarising remarks

1. **Model Extensions** Using the results discovered above, either:
   - Test the effect of including, excluding and combining the parameters on the final model
   - Train a Neural Network model on the above dataset (using e.g. the `TensorFlow` library) and compare results

   [8 marks]
2. **Summary Statement**. Write a short reflective statement (200 words max) summarising a key result from Part 2. You may include a maximum of one figure. Key topics to consider are biases in the dataset, simplified parameters, consequences for future disasaters. A significant fraction of the marks are awarded for reflective thinking: what did you learn? If you did not make it to Part 2, then reflect on your learning from Part 1

   [8 marks]

Additional Marks Marks will be awarded for notebooks, codes and plots that are well explained and well formatted. In particular, attention will be given to sensible variable names, easy to follow comments and notebook structure.

[6 marks]

Maximum possible 22

Total marks available 57
