Summer School 2023 final examination

ME314

Introduction to Data Science and Machine Learning

Suitable for all candidates

Instructions to candidates

This paper has 2 questions, each worth 50 points. You must answer both questions.

This exam is worth *75*% of your total grade.

Deadline: Monday 31st July, 2023, 6pm

ME314 2023 Exam

Instructions

- There are two questions, both worth 50 points each. You should answer **both** questions.
- Complete the assignment by adding your answers directly to the RMarkdown document, knitting the document, and submitting the HTML file to Moodle.
- Please **do not** write your substantive interpretations to the questions in your R comments. They must appear in the knitted HTML document in order for them to receive credit.
- Submit the assignment via Moodle.
- The total word count for this assignment is 1500 words. The word count does not include the code you use to implement the various analyses, but it does include everything else.
- Deadline: Monday 31st July, 6pm

Question 1 – London Cycling Safety

For this question, you will use data on 21492 cycling-involved incidents from 2017 to 2020 in London. These data are stored in the `cycling_severity.csv` file. Your goal is to use this data to build a model to predict the severity of traffic accidents. The data contains the following variables

Table 1: Variables in the `cycling_severity.csv` data.

Variable	Description
`severity_numeric`	A measure of the severity of the incident, ranging from 1 (Not serious) to 10 (Very serious)
`severity_binary`	A binary measure of severity (`"Not Severe"` or `"Severe"`)
`date`	Date of the incident
`weekday`	Day of the incident
`daytime`	Time of day of the incident
`season`	Season of the incident
`weather_conditions`	Weather at time of incident
`light_conditions`	Light conditions at time of incident
`road_surface_conditions`	Road surface conditions at time of incident
`road_type`	Type of road on which incident occurred
`speed_limit`	Speed limit on road
`number_of_vehicles`	Number of vehicles involved in incident
`urban_or_rural_area`	Did the incident take place in a rural or an urban area?
`IMD_Decile`	Index of Multiple Deprivation Decile of area in which incident occurred. (1 means the most deprived and 10 represents the least deprived).
`IncScore`	Income Score (rate) of area in which incident occurred.
`EmpScore`	Employment Score (rate) of area in which incident occurred.
`HDDScore`	Health Deprivation and Disability Score of area in which incident occurred.
`EduScore`	Education, Skills and Training Score of area in which incident occurred.
`CriScore`	Crime Score of area in which incident occurred.
`EnvScore`	Living Environment Score of area in which incident occurred.

Once you have downloaded this file and stored it somewhere sensible, you can load it into R using the following command:

```
cycling <- read.csv("cycling_severity.csv")
```

Your task is to apply at least one of the prediction or classification methods that we covered during the course to this data. You can choose to build a model for predicting either the `severity_numeric` variable (you can treat this as a continuous variable for the purposes of this question) or for the `severity_binary` variable. You can select any model we have covered on the course for this purpose. For instance, you might use a linear regression model, a logistic regression, a random forest model, a ridge regression, and so on.

You will be awarded marks for:

1. Applying your chosen method (15 marks):
    - You should think carefully about which method to apply; which features to include; whether and how to include non-linear relationships; how to select any hyper-parameters of your model; and so on. Although simple models are often powerful, you are unlikely to receive high marks here for implementing a trivially simple model (such as a linear regression with a single explanatory variable).
2. Demonstrating the predictive performance of your chosen method (15 marks).
    - You might, for instance, calculate the MSE of your predictions for the quantitative response, or construct a confusion matrix and calculate accuracy/sensitivity/specificity for a qualitative

response, etc. You will also need to think about the best data to use for evaluating the performance of your model (e.g. training data; train-test split; cross-validation; etc).
- You should provide a comparison of the predictive performance of your model to at least one alternative model specification.

3. Interpreting the result of your method, commenting on how the results might be informative for people working to reduce the severity of cycling accidents (10 marks).
    - You may wish to report some measure of variable importance (plots of fitted values, tables of coefficients, plots of variable importance, etc).
4. Describing what advantages and/or disadvantages your chosen method has over alternative approaches (10 marks).

Your answer should be no longer than 750 words.

Question 2 – NHS Patient Reviews – `nhs_reviews.Rdata`

For this question, you will use a set of 2000 patient reviews of NHS doctors' surgeries across the UK. The data contains the following variables:

Table 2: Variables in the `nhs_reviews` data.

Variable	Description
`review_title`	The title of the patient's review
`review_text`	The text of the patient's review
`star_rating`	The star rating (out of five) that the patient gave
`review_positive`	A categorical indicator equal to `"Positive"` if the patient gave 3 stars or more in their review, and `"Negative"` if they gave 1 or 2 stars
`review_date`	The date of the review
`gp_response`	A categorical variable which measures whether the doctors' surgery provided a response to the patient's review (`"Responded"`) or has not yet provided a response (`"Has not responded"`)

Once you have downloaded this file and stored it somewhere sensible, you can load it into R using the following command:

```
load("nhs_reviews.Rdata")
```

Your task is to apply at least one of the text analysis methods that we covered during the course to this data. Your goal in applying these methods is to generate insight for people who work within the NHS and would like to find ways to improve the service offered by GPs. You can select any text analysis method we covered on the course for this purpose. For instance, you might use a topic model, a dictionary-based approach, supervised classification, and so on.

You will be awarded marks for:

1. Applying your chosen method (15 marks).
    - As with question 1, you should be ambitious here. You are unlikely to receive full credit for running only the very simplest analyses.
2. Discussing the feature-selection decisions you make and how they might affect the outcomes of the analysis (10 marks).
3. Providing some form of validation for your chosen method (15 marks).
4. Interpreting the output of the analysis, commenting on how the results might be informative to people working in the NHS (15 marks).
5. Critically assessing the strengths and weaknesses of your selected approach and proposing at least one alternative text analysis strategy that might be used in the selected application (10 marks).

Your answer should be no longer than 750 words.