Data Hackathon 101

Session 2 - 9th of November

King’s College London Health Science DTC

Schedule

17:00 - Doors open
17:15 - Intro to today's session and Kahoot
17:45 - Machine Learning Models using *scikit-learn*
18:15 - Notebook group work
18:45 - Explaining notebook exercises
19:15 - Introduction to a data challenge and group work

nhanes

National Health and Nutrition Examination Survey

Data:

- NHANES 1999-2000 / 2001-2002
- linked to NDI mortality data

Manual selection:

- only diabetic patients (based on Haemoglobin A1c levels >6.5%)
- potentially relevant variables selected by a clinician (originally intended for survival analysis)

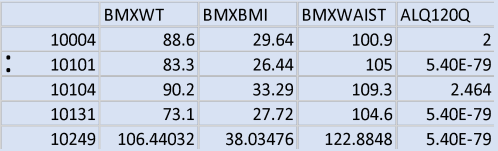
Data source:
https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=1999
https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2001
https://www.cdc.gov/nchs/data-linkage/mortality.htm

Search variables:
https://wwwn.cdc.gov/nchs/nhanes/search/default.aspx
a few variables have been created from raw data, please refer to the lookup file

Problem: Predicting diabetic patient mortality status after 15 years.

Files:

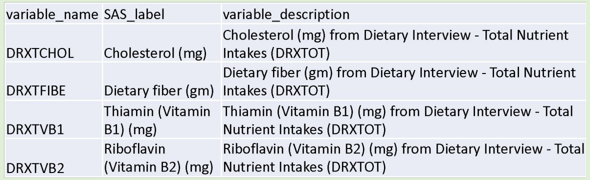
- Main data (missing data imputed):
  - train_data_imputed.csv
  - unseen_data_imputed.csv
- Advanced challenge (optional):
  - train_data_raw.csv
  - unseen_data_raw.csv
- Variable lookup:
  - NHANES_variables_lookup.xlsx



	BMXWT	BMXBMI	BMXWAIST	ALQ120Q
10004	88.6	29.64	100.9	2
10101	83.3	26.44	105	5.40E-79
10104	90.2	33.29	109.3	2.464
10131	73.1	27.72	104.6	5.40E-79
10249	106.44032	38.03476	122.8848	5.40E-79

Variables: 109 variables + mortality status

- demographics
- medical record
- questions about health and habits



variable_name	SAS_label	variable_description
DRXTCHOL	Cholesterol (mg)	Cholesterol (mg) from Dietary Interview - Total Nutrient Intakes (DRXTOT)
DRXTFIBE	Dietary fiber (gm)	Dietary fiber (gm) from Dietary Interview - Total Nutrient Intakes (DRXTOT)
DRXTVB1	Thiamin (Vitamin B1) (mg)	Thiamin (Vitamin B1) (mg) from Dietary Interview - Total Nutrient Intakes (DRXTOT)
DRXTVB2	Riboflavin (Vitamin B2) (mg)	Riboflavin (Vitamin B2) (mg) from Dietary Interview - Total Nutrient Intakes (DRXTOT)

791 observations for training

88 observations for testing

Tasks:

- explain the data to lay audience
- build one or more classifier models to predict mortality status in 2015 (“mortstat”)
- compare and visualise results

Rules:

- use “train_data” only when building model
- “unseen_data” mimics real world scenario when you predict using unseen observations. This is for comparison of multiple models
- If you expose “unseen_data” when training a classifier, this results in “test data leaking” and you’ll get a falsely high performance
- you have complete freedom on feature selection/engineering

Challenge tasks

Challenge:

- Predicting diabetic patient mortality status after 15 years.

Presentation format:

- Powerpoint
- Jupyter Notebook
- No more than 10 min.

Data visualisation:

- How can you visualise data patterns and correlations between variables?

Data Manipulation:

- What exploratory data analysis approaches are used to understand what the data is all about?
- Compare different data manipulation techniques (scaling, normalisation, categorisation). Why would you use any of these?
- What can you do about missing data?

Machine learning:

- Explain the reasoning behind selecting a machine learning model.
- Implement up to three different models (same type, but by varying hyperparameters, or different types). Why certain approaches work better?
- It is possible to combine several models to perform predictions using “ensemble methods”?

..but the most important thing!

There is no need to complete everything!

But show us your thought process of solving a challenge and exemplify the hacking mindset:

- What were the difficulties? How did you solve the particular challenge? What would you have changed or improved?
- What tools have you used? Coding trick? A visualisation or modelling library?