Experience, Portfolio and Suitability

Asset Modelling Data Scientist

Josh Cowley

j.cowley1@ncl.ac.uk

November 17, 2023

Background

Background

- Formally trained as a statistician, Newcastle University

  → MMathStat, Mathematics and Statistics (1st)

  → PhD in Statistics

- Always had an interest in “Computer Science”

- Discovered a career path that incorporates best of both worlds,

  Data Science

Portfolio

Advanced Laryngeal Cancer

Advanced Laryngeal Cancer

Approached by Newcastle Hospitals NHS Foundation Trust.

The pseudonymised dataset contained:

1. survival times (date of diagnosis, death and last check-up)
2. clinical variables (age, smoking status, AJCC tumour status)
3. CT scan output (radiomic data), for a subset of patients

Scope

Objectives were clear from the offset.

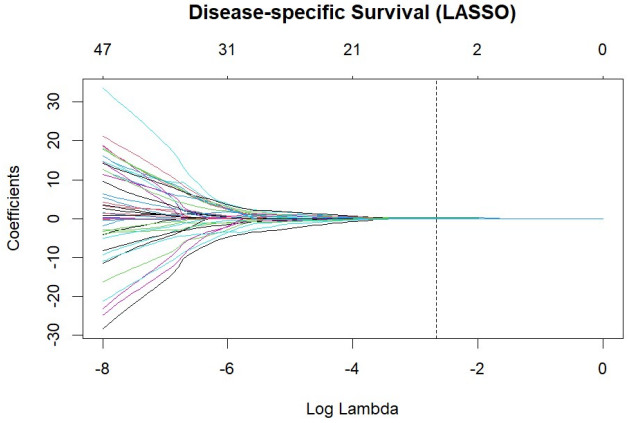
- Quantify impact of clinical predictors on survival times
- Delineate which radiomic features impact survival times
- Assess model improvement when CT data is available
- Publish results in scientific journal [1]

1. Ongoing, currently with *"The Journal of Laryngology and Otology"*.

Approach and Deliverables

Right-censored data, too many radiomic predictors.

Penalised Cox regression, specifically LASSO.



Predictor	Hazard Ratio	CI (95%)	P-value
SHAPE Compacity	2.616	(1.073, 6.378)	0.034 [*]
GLZLM GLNU	1.120	(0.607, 2.066)	0.718
. . .	. . .	. . .	. . .
Age	1.067	(1.012, 1.124)	0.015 [*]
. . .	. . .	. . .	. . .
Treatment (Adjuvant)	0.466	(0.114, 1.908)	0.288

10-fold Cross-validation was used to choose a penalty parameter value.

English Housing Data

English Housing Data

Challenged with developing environmental sustainability indicators; full data science project.

- data collection
- data cleaning
- model building
- time management
- presentation of results
- encouraged reflection

Scope

Investigate environmental health of English housing.

Key questions:

1. Are current metrics (EPC, SAP rating) suitable indicators?
2. Is dwelling efficiency impacted by region within the UK?

Open-source datasets considered: English housing survey (EHS), local authority housing data, and local authority emissions data.

Approach

Intention was always to keep models simple.

Model A

- Linear regression on efficiency or CO2 rating from EPC
- Regressed on key factors like boiler, insulation and more

Model B

- Linear mixed effects model, intercept only
- Models overall mean and region-specific effect

Deliverables

Presented key results to industry partners in May 2022.

Full report and source code available on GitHub.

- Repository (nclJoshCowley/ANNE-challenge)
- Environmental Impact of Housing in England, 2022 (report)

Modelling Hydrocarbon Concentrations in Groundwater Monitoring Networks

Modelling Hydrocarbon Concentrations in Groundwater Monitoring Networks

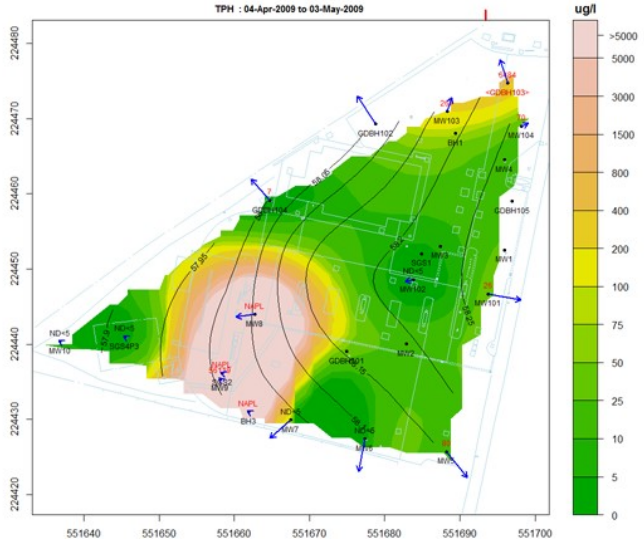
PhD studentship supported by Shell who:

- partially funded research
- provided groundwater monitoring data, specifically hydrocarbon concentrations (subject to NDA)
- communicated suggestions via monthly update meetings

Data

Measurements require extracting sample, transport elsewhere.

- **Spatiotemporal**
  Sampled annually to quarterly at pre-built “wells”
- **Environmental**
  High degree of left-censoring
- **Complex**
  Different geological, chemical, and political considerations



Scope

Two variable “types” arising from shown data generating process:

1. **Analyte**
    - hydrocarbon concentration to be minimised
    - benzene (carcinogenic), toluene, ethylbenzene and more
2. **Predictor**
    - also available and easier data collection
    - temperature, pH, conductivity, dissolved oxygen, ORP

Our intentions were then to

- (*Optimistic*) describe an underlying relationship between analytes and predictors
- (*Realistic*) predict hydrocarbon concentrations using predictors, potentially with telemetry.

> Deeper insight into hydrocarbon concentrations using data analysis on already existing data to increase financial feasiblity.

Approach

A lot of statistical models under supervisor guidance.

- Multiple and multivariate regression [1]
- Tobit regression
- Matrix normal regression
- Mixture of experts
- Varying intercept, spatial prior

1. Dependent variable censored values replaced with half detection limit.

Deliverables

- Very little signal in the data compared to noise
- Aiming to have a letter to the editor, describing findings
- Models to become open source [1], potentially CRAN packages
- More research needed for other water quality variables, say iron
- Advised Shell against exploring telemetry with these predictors

1. https://github.com/nclJoshCowley/bmoe