ACSC/STAT 3740, Predictive Analytics

WINTER 2024

Toby Kenney

Homework Sheet 2

Due: Wednesday 7th February: 11:30

Note: This homework assignment is only valid for WINTER 2024. If you find this homework in a different term, please contact me to find the correct homework sheet.

[Note: all data in this homework are simulated.]

Standard Questions

1. The file `HW2Q1.txt` contains the following data from a company's human resources department about employee retention

Variable	Meaning
`job.title`	The employee's job title.
`job.category`	The type of work.
`salary`	The employee's annual salary.
`age`	The employee's age.
`sex`	The employee's gender.
`experience`	The number of years of experience in the current job.
`training.offered`	Whether the employee was offered training courses.
`training.taken`	Whether the employee took the offered training courses
`retention.5.year`	Whether the employee is still working in the company after 5 years

   Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models. You should take into account any concerns with data collection and processing.

2. The file `HW2Q2.txt` contains the following data about school performances in standardised tests for Grade 8:

Variable	Meaning
no.students	The number of students in Grade 8 attending the school.
teacher.student.ratio	The average number of students per teacher in a class at the school.
funding	The schools source of funding — government, independent or private.
specialist.teacher	Whether the school employs teachers with specialist knowledge for each subject.
teacher.5.years	The percentage of teachers at the school with at least 5 years of experience.
parent.employment	The percentage of parents of children at the school who are employed.
median.parent.salary	The median salary of parents of children at the school
mean.parent.education	The average number of years of full-time education of parents of children at the school.
average.score.mathematics	The average score of children in Grade 8 at the school on the standardised mathematics test.
average.score.english	The average score of children in Grade 8 at the school on the standardised English test.

The test results were published by the examination board. Information on schools was provided by the schools. Information about parents of children was taken from surveys conducted by the schools.

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models. You should take into account any concerns with data collection and processing.

3. The file `HW2Q3.txt` contains the following data from an experiment about survival times under a certain treatment for a disease.

Variable	Meaning
severity	The severity of disease symptoms prior to treatment 1=mild, 10=very severe.
treatment.wait	The number of days the patient needs to wait before being treated.
age	The patient's age at first diagnosis.
sex	The patient's sex.
health.index	An index assessing the patient's overall health — 100=perfect health, 0=dead
outcome	Whether the patient dies before the end of the study period.
time.to.outcome	The time in years before the patient dies or the end of the study, whichever happens first.

Severity is assessed by the treating physician at first consultation. Treatment wait is calculated as the difference between the date of the first consultation and the start of treatment, from the hospital database. Age and sex are from the patient's medical records. Health index is based on physician assessment and a questionaire filled out by the patient. Outcome and time to outcome are from the hospital records.

4. A credit card company is improving its fraud detection models. It collects the following data for all purchases made:

Variable	meaning
date	The date of the transaction.
time	The time of the transaction.
location	The location of the transaction.
last.location	The location of the previous transaction from this credit card.
online	Whether the transaction was online.
account.balance	The amount owing on the card before the transaction.
credit.limit	The credit limit of the card.
recent.spending	The amount spent on the card in the previous week.
purchase.amount	The cost of the purchase.
fraudulent	Whether the purchase was fraudulent.

Most of the data are automatically processed at time of purchase. Fraudulent is based on customer reports and manual reviews. The data are in the file `HW2Q4.txt`.

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models.

5. A scientist is studying the effect of climate on road maintainance costs. She collects the following data from a large number of cities

Variable name	Meaning
ave.winter.temp	The average daily maximum temperature for the months January–March.
ave.summer.temp	The average daily maximum temperature for the months June–August.
total.precipitation	The total yearly rainfall (mm)
days.below.freezing	The number of days during the year where the maximum temperature is below $0°C$.
total.snow	The total annual snowfall (cm)
car.usage	The total distance driven annually by all inhabitants of the city. (km)
annual.costs	The total annual costs for road maintainance ($).

The data are in the file `HW2Q5.txt`. The weather data are from historical weather records at nearby weather stations. Car usage is estimated from surveys. Annual costs are from local government reports.

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models.