ACSC/STAT 3740, Predictive Analytics

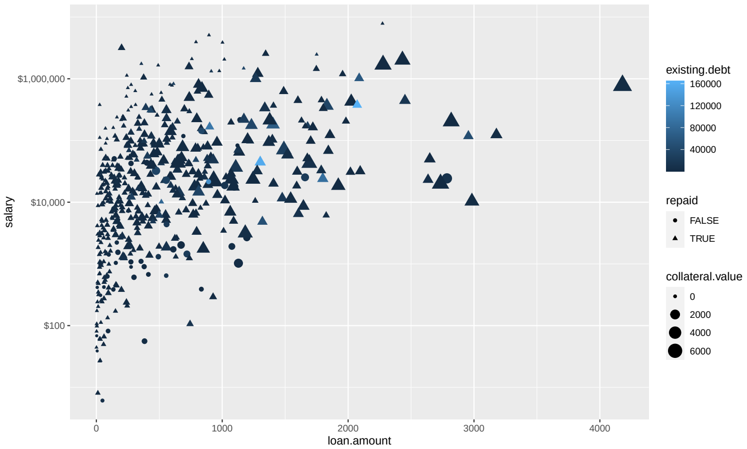
WINTER 2024
Toby Kenney

Practice Final Examination

This Sample examination has more questions than the actual final, in order to cover a wider range of questions. Estimated times are provided after each question to help your preparation.

[Note: All data on this exam are simulated.]

1. Use `ggplot` to produce the following plot from the data in file `PFQ1.txt`.



2. The file `PFQ2.txt` contains the following data from an experiment to determine the effect of fertilisers on growth of bean crops:

Variable	Meaning
Soil.type	The type of soil
Water	The amount of water (l/hectare) provided to the crop
Species	The species of bean
Planting.date	The date on which the crop was planted
Fertiliser.A	The quantity of fertiliser A (l/hectare) provided to the crop
Fertiliser.B	The quantity of fertiliser B (l/hectare) provided to the crop
Yield	The quantity of crop harvested (kg/hectare)

Construct a plot or plots to show this data for the purpose of data exploration.

3. The file `PFQ3.txt` contains the following data on advertising campaigns from a company's marketing department.

Variable	Meaning
Type	The medium of the advertising campaign
Duration	The length of time the campaign was running
Targeted	Extent to which the campaign was targetted on a scale 1–5
Cost	The total amount spent on the advertising campaign.
Target.audience	The number of individuals expected to see the advert
Sales.Increase	The increase in total sales in the two-month period after the campaign, compared with the two-month period before.

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.

4. The file `PFQ4.txt` contains the following data from an experiment about pollution and bacteria. The data set contains the following variables.

Variable	Meaning
NO2	The concentration of nitrogen dioxide in the water (ppm)
SO2	The concentration of sulphur dioxide in the water (ppm)
pH	The alkalinity of the water (7 is neutral, 0 is strong acid, 14 is strong alkali)
Temp	The temperature of the water ($^\circ C$)
Cyanobacteria	The abundance of Cyanobacteria in the water
Firmicutes	The abundance of Firmicutes in the water
Actinobacteria	The abundance of Actinobacteria in the water

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.

5. The file `PFQ5.txt` contains the following data from a company's quality control department. The company is monitoring the number of defective products produced over time and wants to develop a method for quickly detecting when the machine has started to produce defective units.

Variable	Meaning
Machine	The indentification number of the machine
Batch.no	The number of the batch produced. Machine batches of 1,000 units are produced sequentially, starting from number 1.
Production.time	The time taken to produce the batch
Power	The amount of energy used by the machine
Defective	The number of defective units in the batch

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.

6. A doctor has collected the following data about age and various health conditions in the file `PFQ6.txt`.

Variable	Meaning
Age	The age of the patient
BMI	The BMI of the patient (weight (kg)/(height (m))^2
SBP	Systolic blood pressure of the patient
Respire	Respiratory rate of patient (breaths/second)
Heart	The heart rate of the patient
Glucose	Blood sugar of patient
RBC	Concentration of red blood cells
WBC	Concentration of white blood cells

Fit a random forest model to predict the number of white blood cells for a patient from the other predictors. Use this model to predict the concentration of white blood cells for the patients in the file `PFQ6_test.txt`.

7. An insurance company has collected the following data on fire insurance claims in the file `PFQ7.txt`.

Variable	Meaning
doy	Day of year 1=1st January, 365=31st December
rent	The monthly rent for the premises
alarm	Whether the premises has a fire alarm
sprinklers	Number of sprinklers on the premises
area	Total area of the premises
claim.amount	The amount claimed for the loss event

Fit a generalised linear model, with a gamma response variable, using a log link, and log-transforming the predictors rent and area, to predict the claim amount from the other predictors.

Use this model to predict the claim amounts for the claims in the file `PFQ7_test.txt`.

8. A scientist has collected the following data about planets in the file `PFQ8.txt`.

Variable	Meaning
Size	The radius of the planet (km)
Mass	The mass of the planet (tonnes)
Composition	What the planet is mostly made from
Star.size	The radius of the star that the planet orbits
Orbit.distance	The distance at which the planet orbits the star
Revolution.time	The time taken for the planet to complete a turn on its axis
Earth.distance	Distance from Earth's solar system (light-years)
Oxygen	Whether the planets atmosphere has a detectable quantity of oxygen.

Fit a generalised additive model to determine whether a planets atmosphere has a detectable quantity of oxygen, and use it to predict the results for the planets listed in the file `PFQ8_test.txt`.

9. A data scientist is analysing data about spending and financial security in the file `PFQ9.txt`.

Variable	Meaning
Year	The year in which the data were collected
Family.size	The number of individuals supported in the household
Food.spending	Average Total monthly amount spent on food
Clothing.spending	Average Total monthly amount spent on food
Housing.spending	Average monthly spending on rent or mortgage payments
Travel.spending	Average monthly spending on travel
Entertainment.spending	Average monthly spending on entertainment
Annual.salary	The annual combined salary of the household
Pension.percent	The expected annual combined pension of the household, expressed as percent of current income
Age	The age of the highest earner in the household
Debt	The total level of debt owed by the household
Assets	The total value of assets owned by the household
Retirment.success	Whether the family was able to retire at their planned time or earlier with the intended pension

He has fitted a linear model to predict which households will successfully be able to retire, using the code in the file `PFQ9.R`. Perform diagnostics to test which of the assumptions of this model are reasonable. What changes would you suggest making to the model to better model the data?

10. An actuary is reviewing data about catastrophe insurance claims in the file `PFQ10.txt`.

Variable	Meaning
Event.type	The type of catastrophe event
Area.size	The size of the area affected (km^2)
Area.people	The number of people living in the affected area
Total.damage	The total claims for the affected area

She has fitted a generalised additive model, a random forest model and a generalised linear model including a number of interaction terms and polynomial terms, to predict the total damage, using the code in the file `PFQ10.R`. Assess which of these models is better at predicting the data. [You may need to modify the code provided to do this.]

11. A scientist has analysed some data and written the following conclusion to their paper.

The purpose of this analysis was to determine the extent to which an individual's fertility can be predicted from genetic data. The data were taken from the study of [1]. The researchers collected genetic samples from childless couples, both aged 20–30, who were trying to conceive. They then followed the couples over a 1-year period, recording which of them successfully conceived during that period. They also conducted a survey about other factors that may be influencing their success.

There were a total of 32,041 couples enrolled in the study, from four different cities: London, UK; Montreal, Canada; Qingdao, China; and Rome, Italy. The city where the couple were based was included as a predictor variable, since the effects of several predictor variables could be affected by the location. The study was conducted over a period of 3 years from 2013–2016. There was an attrition rate of 13% in the study. Couples who did not complete the study were removed from the analysis (even if they had conceived before that time). This approach is open to some concerns, since couples who stopped the study might be different from couples who remained in the study. It is possible that more couples who failed to conceive left the study, leading to potentially biased results. Further research is needed to develop a better method to handle attrition. Alternatively, a study with a shorter follow-up period and more couples might have lower attrition rates, leading to less potential bias in the conclusions.

The genetic data comprised 12,358 single nucleotide polymorphisms (SNPs) from each male participant in the study, and 12,704 SNPs from each female participant. Thus for each couple, there are a total of 25,062 genetic predictors, in addition to 72 other predictors from the survey. There were some missing responses from the survey variables, and other ambiguous responses were removed from the data. [1] removed three of the survey predictors from the analysis — sex time of day, menstual frequency, and sex frequency, because many responses were missing for these predictors. In our analysis, we were able to include these responses because the random forest method we used is able to handle missing values.

[1] was unable to find any significant genes associated with fertility when correcting for the effect of other predictors. However, the methodology used was very conservative. In this paper, we used a new approach to the statistical modelling, which is more able to identify interactions between genes. This approach was first suggested by [2]. The basic idea is to divide the data into two parts, use the first part to screen the genetic predictors, to find a shortlist of the predictors most likely to be associated with fertility, then to use the second part of the data to fit a random forest model to predict number of children from the selected predictors.

The work of [2] and [3] has shown that this approach can be very effective at identifying complex interactions between genes in other contexts. We have modified the approach slightly, following the suggestion of [4] to divide into two subsets for screening the predictors. We found that this approach produced a lower test error on the data.

To assess the predictive performance of our method, we performed our approach using two thirds of the data as training data, and one third of the data as test data. The training data was divided into two screening subsets, each comprising one sixth of the training data, and one modelling subset comprising the remaining two thirds of the training data. Cross-validation was used on the modelling subset to select the tuning parameter for random forest. 81% of the couples who completed the study successfully conceived during that period. We found that using only the survey predictors, we were able to predict with 86% accuracy whether a couple would conceive. Using the genetic data, we were able to increase this to 88%. This clearly indicates that the genetic data is useful for predicting conception success.

For comparison, we used LASSO to fit a generalised linear model to predict whether a couple would conceive. For this model, there was no improvement in test accuracy over the model with just the survey predictors. This indicates that the effect of the genetic data on fertility is not linear. Since the genetic variables are mostly binary, this indicates that interactions between the genes are responsible for the effect on fertility.

From the variable importance given by random forest, the most important genes are MEB12, INT21 and RMS7. Fitting a random forest model using only those three genes produces a test accuracy of 87%, indicating that the effect of genes on fertility is spread over a large number of genes, each of which has only a minor effect. The LASSO fitting with logistic regression selected 4 genes: MEB12, RMS7, OLB13 and QJA14. The first two of these were important in the random forest method. However, INT21, which was important under random forest was not selected by LASSO, suggesting that this gene is only important because of its interaction with other genes. The genes OLB13 and QJA14 were very low in the random forest variable importance. This may suggest that these genes are surrogates for a combination of other genes, that may be better predictors in a nonlinear model.

We also looked at different performance measures. In particular, we considered weighted accuracy, where we increased the weight of the couples who failed to conceive, so that both groups had the same weight. Under this performance measure, using only the survey variables, the LASSO logistic regression achieves a weighted test accuracy of 64%, while random forest achieves a weighted test accuracy of 71%. When the genetic variables are included, the weighted accuracy improved to 66% for LASSO and 74% for random forest. This shows that the genetic predictors are important even in the linear model.

write an abstract for the paper.

12. The following quotes come from a report on the effect of financial hedging on a company's profit. Where in the report should they be placed? Justify your answers.

    (i)

The MSE of the random forest model on the test data was 22.4. This is not significantly better than the generalised additive model. Because the generalised additive model is more interpretable, we therefore prefer it to the random forest model.

(ii)

While hedging against adverse financial conditions reduces the expected profits by 2%, it greatly reduces the probability of a large loss. Given the company's limited capacity to absorb large losses, we consider this reduction in risk to justify the proposed hedging scheme.

(iii)

Previous work by [1] used a neural network to predict profit values. While the accuracy of their predicted profits was reasonable, their method ignored the dangers of currency fluctuations and the timing of conversion. [2] modified this approach to account for timing of currency conversion. However, their changes to the model caused a decrease in accuracy on the test data.

(iv)

There may be potential to develop a hedging scheme that is able to achieve a similar reduction in risk with lower reduction in profit, for example, the schemes suggested by [3] are promising. However the proposed scheme has a major advantage of simplicity, which can lead to reduced implementation costs. The reduction in implementation costs, which can be quite substantial [4], should be enough to outweigh any theoretical advantages of the superior hedging scheme.

13. A scientist has analysed the data in the file `PFQ13_a.txt` using the commands in `PFQ13.R`. The data show the change in log abundance of a number of common gut bacterial genera, in response to treatment of patients with antibiotics. For reference, the taxonomy of the relevant bacteria is in the file `PFQ13_b.txt`. She has concluded the following:

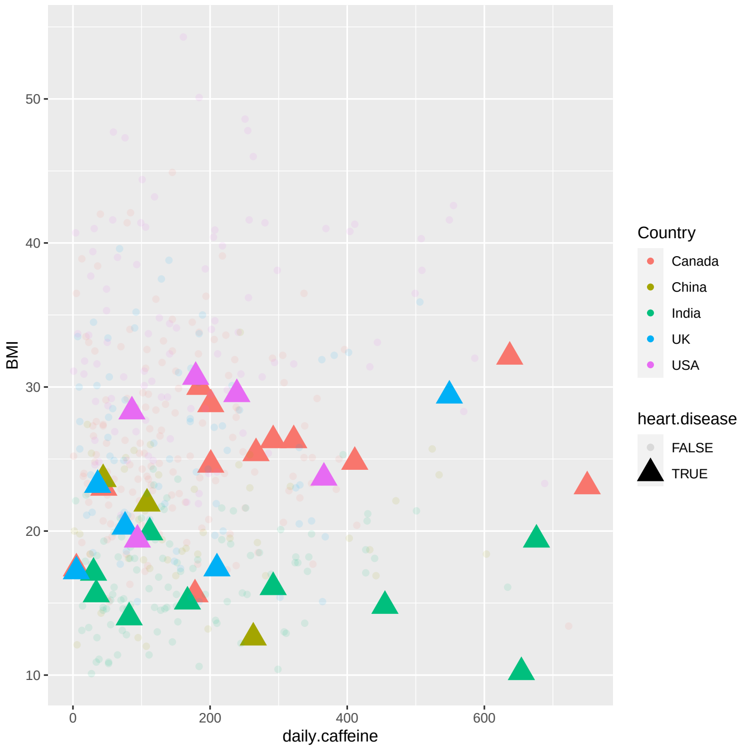
    (a) The phyla *Firmicutes* and *Proteobacteria* have very uniform levels of reduction in response to glycopeptides.

    (b) The classes in phylum *Firmicutes* react in very different ways to tetracyclines

    (c) From the responses of 6 indicator genera — Actinomyces, Akkermansia, Clostridium, Desulfovibrio, Fusobacterium, and Roseburia — it is possible to predict the response of all other genera with reasonable accuracy.

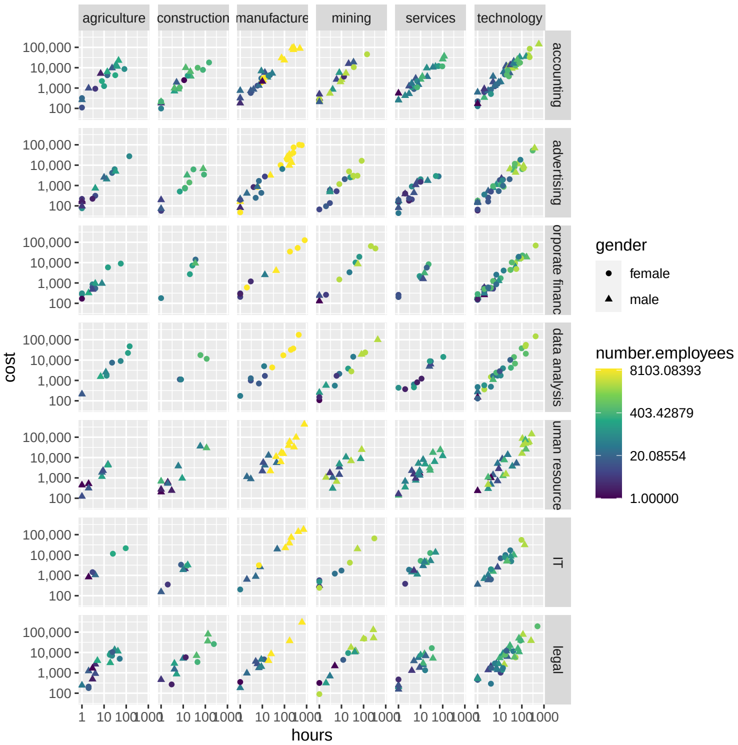
    Display the data and analysis results so as to demonstrate the conclusions.

14. The file `PFQ14.txt` contains data from an experiment about the relation between caffeine and heart disease. The data are not formatted in a very convenient way. Read the data into `R` and reformat into a more convenient way, and use it to create the following plot.



Make a list of all corrections made to the data.

15. The file `PFQ15.txt` contains data from a consultancy company about customers and services. The data are not formatted in a very convenient way. Read the data into `R` and reformat into a more convenient way, and use it to create the following plot.



Make a list of all corrections made to the data.