ACSC/STAT 3740, Predictive Analytics

WINTER 2024

Toby Kenney

Homework Sheet 5

Model Solutions

Standard Questions

1. *A scientist is studying the factors affecting educational performance of students aged 16–18, and has written a paper with the following conclusion section.*

Educational achievement is essential in the modern world, with the majority of jobs requiring at least an undergraduate education [1], and some sources predicting that by 2035, over 80% of jobs will require this level of education [2].

Given this environment, ensuring educational success of school-age children is imperative. However, many children are not achieving this success. In order to prepare the next generation for the challenges they will face, we must learn more about the factors that contribute to success.

Previous research [3] in this area has suggested that school hours less than 4 is detrimental to performance, and that family education level is highly correlated with success, particularly in developed countries.

We analysed data from 1520 schools in 37 countries. For each school, we collected data on the resources available to the school, the students' backgrounds, and the amount of time students dedicate to study. Performances of schools were assessed based on the average results of standard international PISA test scores taken by all students at the school.

Our dataset is more comprehensive than previous research [3,4] in that it includes potentially relevant predictors, "teacher training years" and "student tutoring access" which have been neglected in previous studies.

We excluded 4 schools from our analysis, because the surveys used to collect data on parent education levels and student use of tutoring services had non-standard variations of the questions, so it was unclear how the values relate to the corresponding values at other schools. We also consolidated the data on parent education into 4 levels, in order to make the results comparable across different countries.

We also calculated additional features from the available data. We calculated the resources per hour of study by dividing the total resources per student by the number of hours of study. We calculated classroom proportion as the area of all classrooms divided by

the school's total area. These predictors have been suggested as possible predictors in the literature [5], but that paper did not have the necessary data to calculate them.

We used two methods to analyse the data. The first method was a random forest predictor. The second was a generalised additive model. Previous research [4,5] used linear regression to estimate PISA scores. However, this model does not work well when the relation between predictors and school performances is nonlinear, and many of their conclusions were counterintuitive.

We compared the performances of the two models using cross-validated MSE. We found that the cross-validated MSE was 12.4 using random forest and 14.1 using the generalised additive model. Because interpretability is critical to this project, we prefer the generalised additive model.

The most important predictors were "log school funding per study hour", which had a linear effect; "parental education", which was treated as a categorical variable; "school hours", which had a non-linear effect; "student tutoring access", which had a linear effect; and "teacher salary", which had a non-linear effect. These variables were also ranked highly by the random forest. However, the random forest also ranked "family status" highly, despite the GAM fitting the effect of this predictor as almost zero. This could indicate that this variable is important mainly because of its interaction with other predictors.

The selected variables mostly agreed with previous literature. However "teacher salary" had not previously been identified as such an important predictor. This may be because of the high correlation with "log school funding" and non-linearity of the effect.

From a Q-Q plot, we see that the residuals from the GAM are approximately normal. We also see that the residuals are unrelated to the fitted values, suggesting that the assumptions of homoskedacity and normality are reasonable. There are a few influential points, but the residuals at these points are small, indicating that the points follow the general trends fitted by the model.

The data were the largest dataset studied on this subject to-date. Unfortunately, the data involved combining data from a large number of sources, and it was not possible to guarantee that the exact classifications used by the different sources were identical. Some predictors were not truly equivalent because of economic differences between different countries. However, sub-group analyses in each country did confirm similar trends in the majority of cases.

We performed sensitivity analyses to determine the extent to which differences in converting variables between countries would affect the results of the analyses. Details of these analyses are in Appendix 2. These analyses indicate that using a different conversion method between different countries would not significantly change the results of the analysis.

The GAM improves upon previous research by allowing non-linear effects, and produces interpretable results that are in line with previous literature. However, the additive structure of the model can fail to model interactions between predictions. This could explain why random forest achieved a lower mean squared error.

For future research, it is important to improve the data collection to ensure that data are comparable for schools in all countries. Many of the variables are collected by the schools and local governments. However, many of these schools and local governments collect additional data that could be used to reduce the inconsistencies between different regions in the current data set.

Another important topic for future research is to refit the GAM with suitable interaction terms between the predictors. This will be a more flexible model that is able to model more complicated functions of more than one variable.

write a 150-word abstract for this paper.

The purpose of this study was to determine the key factors affecting education performance of schoolchildren aged 16–18. Performance results on standardised international tests were collected from 1520 schools in 37 countries. Data were also collected on school resources, students' background and study time.

The data were analysed using random forest and GAMs. The methods give similar MSE, with random forest slightly better at prediction. We prefer the GAM for its interpretability. The most important predictors were linear effects in "log school funding per study hour" and "student tutoring access"; non-linear effects in "school hours" and "teacher salary", and the categorical variable "parental education". These are mostly inline with previous findings, but the imporance of "teacher salary" has not been previously reported.

In future we hope to standardise the data collection to ensure more comparable results across countries, and to model key interactions between predictors in our GAM.

2. *The following quotes come from a report on the pros and cons of stock options as a means of employee compensation. Where in the report should they be placed? Justify your answers.*

   (i)

In conclusion, stock options are invariably a less effective means of employee remuneration for employees on salaries under $100,000. For employees with leadership roles and salaries above $100,000, stock options are often a valuable part of a compensation package, and can be mutually beneficial, depending on a number of factors.

This is a very concise take-away message, and as such, should probably be in the **executive summary**. Given the interpretation involved in the statement, it would not be in the results section, where

we would expect to see a more precise statement of how stock options are less effective. In the conclusions section, we would expect a little more depth. However, this could be the start of a paragraph in the conclusion, which gives more depth.

(ii)

The main purpose of using stock options as reimbursement is to give employees an incentive in the company's success. Many sources consider this an ingenious method for extracting optimal performance from employees [1,2]. However, other authors have questioned the practice, suggesting that the benefits are overestimated [3], and the costs are underestimated [4].

This is probably from the **Introduction**. It is providing extensive and detailed background to the problem. It is probably too detailed for the executive summary, particularly since the target audience for such a report would probably be expected to be familiar with a lot of this background. If the report were targeted at an audience that was not expected to be familiar with the use of stock options for employee reimbursement, then these statements could be included in the executive summary, but would usually be condensed.

(iii)

We fitted three models to predict employee recruitment and retention from the available predictors. The first model was a logistic regression model; the second was a logistic regression with transformed predictors; and the third was random forest. The performances of these models on test data are shown in Table 5. We see that the random forest model had lower misclassification error, but lower log-likelihood than the logistic regression with transformed predictors.

This is clearly from the **Model Selection and Interpretation** or **Results** section. It describes and compares the outcomes of the analysis in some detail, but not sufficient detail that it would be moved to an appendix.

(iv)

We see that "employee salary" is heavy-tailed, and highly correlated with "options awarded". Based on this, a natural approach is to log-transform employee salary and calculate options awarded as a percentage of total remuneration.

This is from the **Data Exploration** section. It describes a preliminary examination of the data, and the transformations done prior

to analysis.

(v)

The analysis uses options as a proportion of total remuneration as a predictor. We observed that there is a valid concern that this depends too much on the valuation methods used for various parts of the remuneration. The methods used for the valuation are based on the accounting practices actually used by the company and represent the actual reduction in reported profits due to these remunerations. However, we want to know how sensitive our results are to these valuation methods. It is possible that the efficiency of these remuneration methods depends on their perceived value, which may be based on a different valuation method. We therefore compared the analysis results using the different valuation schemes described below.

This is probably from the **Appendix**. It is going into significant detail about a minor point in the data analysis. If it were deemed sufficiently important to be included in the main report, it would probably be in the **Results** section.

(vi)

We preferred a logistic regression model as the best compromise between prediction and interpretability. This model selected a number of interaction terms involving "option remuneration percentage", most significantly the interaction with "salary" and with "employee hours".

This might be from the **Executive Summary**. It is a fairly succinct description of the selected model and its main conclusions. It may be too technical for the executive summary, and might also appear in the **Results** section.

(vii)

The conclusions of all three methods were relatively similar in terms of the important predictors. Random forest ranked the predictor "supervisor management style" as an important predictor, while this predictor was not significant under the logistic regression models. This suggests that this predictor may be important because of some unmodelled interactions with other predictors. This is consistent with the previous results in the literature [5,7,9] which modelled particular interactions of this predictor with other predictors.

This is clearly discussion of the analysis results and their limitations, and as such belongs in the **Conclusion** section. Some parts of this quote could be in the results section, but comparing detailed results

with the literature is more appropriately put in the Conclusion section.

(viii)

The calibration plots showed that the logistic regression model with transformed predictors was better calibrated than the other models, but was not perfectly calibrated, suggesting that some of the predictors' effects are non-linear, or that there are unmodelled interactions, or confounding variables that are not included in the data set. Variables relating to the background of the employees are an obvious omission from this dataset.

This is discussing the limitations of the analysis, and would usually be put in the **Conclusion** section.

3. *A data scientist has analysed the data in the file* `HW5Q3.txt`*. The data show the relation between lifespan and rate of evolution.*

Variable	*Meaning*
Origin	*The estimated time of the last common ancestor of this genus.*
No.species	*The number of species sequenced from this genus.*
Evolutionary.distance	*The average evolutionary dissimilarity between a pair of species from this genus*
Genome.length	*The average length of a genome from this genus.*
Lifespan	*The average lifespan of an organism from this genus.*
Diet	*The animal's diet — carnivore, herbivore, ominvore, insectivore*
Ave.litter	*The average litter size of the animal (the number of offspring produced in a single pregnancy.*

She has concluded the following:

(a) *The evolutionary distance is positively associated with origin time and average litter size. It is negatively associated with lifespan.*

(b) *The variance of evolutionary distance decreases with genome length.*

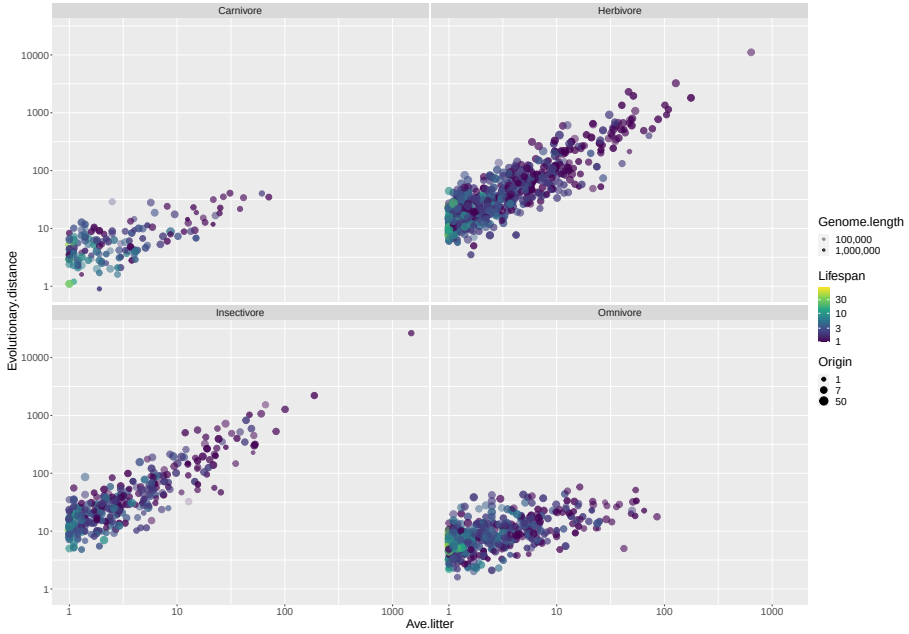
(c) *Herbivores and insectivores evolve faster on average, but litter size has a much stronger association with evolutionary distance for these animals.*

(d) *The variance in evolutionary distance decreases with number of species sampled, but only by a small amount.*

Display the data so as to demonstrate as many of these conclusions as possible.

This is a challenging plot to create because we want to show the interaction between a large number of variables. Since we want to show that shorter genomes have higher variance, we can map alpha to log genome length.

This way, the more opaque points will be in the centre of the trend, and transparent points on the outside. The most obvious way to represent diet is a facet wrap. Mapping x to average litter size makes sense, as it makes it easy to compare different trends across facets. This leaves size for time since origin and colour for lifespan, which just about allow the patterns to be discerned. The predictors are all heavy-tailed, so a log transformation is probably appropriate for all predictors.



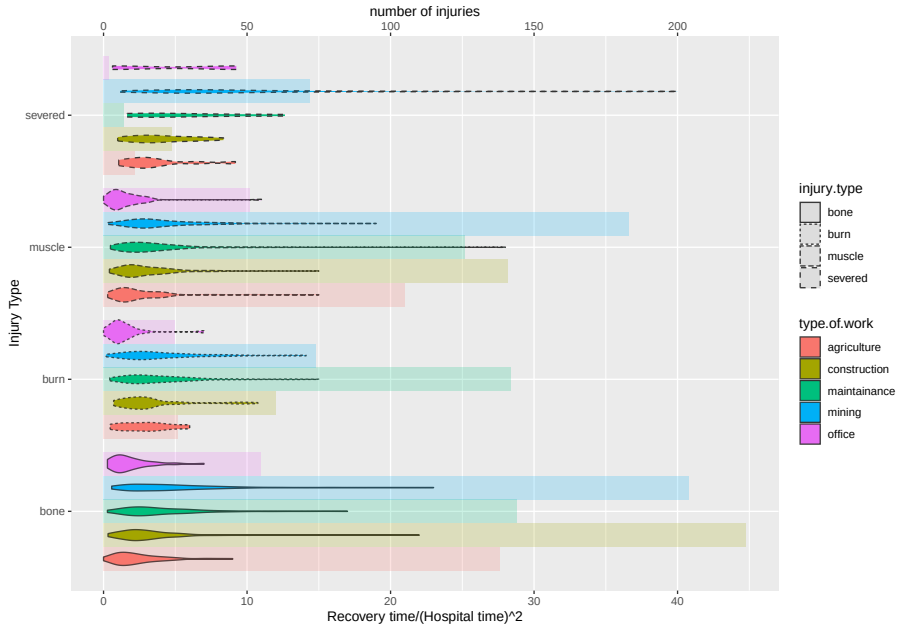
It was produced using the code

```
HW5Q3<-read.table("HW5Q3.txt",stringsAsFactors=TRUE)
library(GGally)
library(dplyr)


ggplot(HW5Q3,mapping=aes(y=Evolutionary.distance,
                         size=Origin,
                         x=Ave.litter,
                         alpha=Genome.length,
                         colour=Lifespan))+
    geom_point()+
    facet_wrap(~Diet)+
    scale_colour_viridis_c(trans="log",breaks=c(1,3,10,30))+
    scale_size_continuous(trans="log",breaks=c(1,7,50))+
    scale_y_log10()+
    scale_x_log10()+
    scale_alpha_continuous(trans="log",breaks=c(10000,100000,1000000),
                           labels=scales::comma)
```

This plot does not show number of species. We could use a facet grid or shape to show the number of species grouped by bins, but it is really hard to show the pattern from this. An option would be to show residuals of the fitted regression model coloured by number of species, but this would require additional details of the models used.

4. *A data scientist has analysed the data in the file* `HW5Q4.txt`*, and produced the following plot of the results. The data are from a workers' compensation insurance company, and the purpose of the study is to improve prediction of the length of payments for disability benefits.*



Variable	*Meaning*
age	*The age of the injured employee*
sex	*The sex of the injured employee*
type.of.work	*The type of work the employee does*
injured.part	*The body part that was injured*
injury.type	*The type of injury*
hospital.time	*The number of days the employee spent in hospital.*
recovery.time	*The number of days before the employee was able to return to work.*

Write a paragraph to describe the figure and the conclusions drawn from it.

[In case the figure is not clear, the bars show the counts of injuries and the violin plots show the distribution of recovery time over squared hospital time.]

From Figure 1, we see that the most common injuries are bone and muscle injuries, and the least common injuries are severed body parts, which are rare for all types of work with the exception of mining. Mining and construction are the industries which produce the most bone and muscle injuries, while maintainance produces the most burns. Office work produces fewest injuries of all types. For each injury type, office workers are the quickest to return to work, followed by agriculture workers. Severed body parts have the longest delays in returning to work, with the other injury types having comparable recovery times for a given hospital stay.