ACSC/STAT 3740, Predictive Analytics

WINTER 2023

Toby Kenney

Homework Sheet 5

Model Solutions

Standard Questions

1. *A data scientist is studying the effectiveness of advertising campaigns and has written the following conclusion to her report.*

Our annual advertising budget is $2.2 million. This represents a substantial investment, and it is essential that we identify the most effective way to spend this money. To study this problem, we examined previous advertising campaigns run by ourselves and our competitors. For campaigns run by our competitors, we asked our advertising consultants to estimate the costs of each component of the campaign (the campaigns were divided into six components: online targetted, online general, TV targetted, TV general, Other targetted, and other general). The "other" components included advertising via radio, newspaper and billboards. The sales in the periods preceding and following the campaign were found from the annual financial reports of the companies.

In addition to the estimated costs of various components of the advertising campaign, and measures of the company's success, our data set also included details about the company (market capitalisation, headquarters location, years in business, reputation, line of business), the product (quality, style, weight and customisation) and the type of advertising campaign (general, new product, promotional deal).

The estimates from our consultants represent a very cursory examination of the cost to run the campaign, and are based on plausible assumptions about the amount of time planning or designing the campaign. However, it is possible that these represent significant misestimates of the actual costs of the campaigns, resulting in our fitted model being inaccurate.

A preliminary examination of the data indicated that the data from several campaigns was unreliable. We therefore removed these campaigns from the data. These campaigns were mostly particularly unusual, and often from smaller companies. Thus, their removal may lead to sampling bias in the analysis results.

We compared two methods for predicting the increase in sales due to the advertising campaign. The first was a random forest, which is a very flexible method, but less interpretable. The second was a generalised additive model, which gives a nice interpretable model, but does not model interactions between predictors. We fitted these models using both a "sales gain" response variable, which was the total increase in sales from the period

preceding the campaign and a "log sales gain" response, where we calculated the logarithm of the ratio between the sales for the period during and after the campaign and the sales for the period before the campaign. The log-transformed response variable was preferable for both methods, both in terms of improved prediction (evaluated using cross-validated MSE with 10 folds on both the log ratio and sales difference scales) and qualitatively in that the residuals were more normal and more homoskedastic after the log-transformation.

Based on the Random forest model, it appears that the interaction between product quality and company reputation is an important predictor. This is consistent with experience, as customers are only willing to pay higher prices for quality products if the company's reputation is sufficient to support the claimed quality of the product. We therefore added this interaction term to the GAM model and saw a large improvement in the cross-validated MSE.

Even after adding the interaction terms to the GAM model, Random Forest still produces the best prediction in terms of cross-validated MSE. However the differrence is not significant (10.52 with standard deviation 0.63 versus 11.04 with standard deviation 0.45 on the log-scale). Since the purpose of this analysis is to develop a good advertising strategy for our own campaigns, the interpretability of GAM is extremely valuable, since it allows us to more easily predict what changes to an advertising campaign would increase profitability.

Since the number of predictors was not too large (15 predictors with 490 observations), we found that variable selection in the GAM model was not required and did not improve prediction. All of the predictors were significantly associated with the sales gain under the GAM.

Based on the fitted GAM, the most important predictors are online targetted, type of campaign and product quality. For the log-scale response, the effect of online targetted spending is non-linear, showing the expected diminishing returns for additional advertising spending. Looking at the fitted spline functions for each type of spending, it seems that targetted online spending is the closest type of spending to having a linear relation with sales increase. The fitted relation between product quality and sales increase is also nonlinear, with larger sales increases for both high quality and low quality goods. This is also consistent with experience, where sales of both high-quality and economy goods can be improved by advertising.

Using the fitted model, we estimate that the most effective strategy for our upcoming advertising campaign would be to spend \$1.2 million on targetted online advertising; \$700,000 on general Television advertising; \$200,000 on targetted Television advertising; and \$100,000 on general other advertising. This strategy is successful for all model parameters in a confidence region about the estimated parameters. We also analysed this strategy under the fitted Random Forest predictor, and found that Random Forest also predicts strong returns from this strategy.

We found confidence intervals for the estimated model parameters. Based on these confidence regions, the best strategy for our campaign might be spending between \$800,000 and \$1.5 million on targetted online advertising; between \$0 and \$200,000 on general online advertising; between \$600,000 and \$800,000 on general television advertising; between

$100,000 and $600,000 on targetted television advertising; and between $0 and $300,000 on other advertising.

There are a few potential sources of error in this analysis. Firstly, the data was based on estimates of previous campaigns, and it is possible that these estimates are inaccurate, which may create bias to our model. We performed some simple influence analysis, where we changed some of the key assumptions of our analysis. We found that most changes had relatively small effect on our conclusions. However, systematic underestimation of click rates for targetted online advertisements could lead to a substantial change in the estimated optimal strategy. Unfortunately, there is no obvious way to mitigate this risk.

The second potential source of error is in the assumption that the effectiveness of particular advertising campaigns will remain the same. The data was based on advertising campaigns between 2010 and 2019. We did not detect trends in the data, but it is possible that different types of advertising campaign will be perceived differently in the future, leading to our model's predictions being inaccurate.

The third potential source of error is in sampling bias in our choice of competitor companies. The choice of competitor companies was based on three factors: 1) Companies operating in at least three countries. 2) At least 30% of the companies revenue is in the same industry as our company, and the advertising campaign focused on products in this industry. 3) Company's headquarters in Canada or USA. This caused us to omit a number of companies that could be argued to be closely related to our company, and to include a number of companies that many would not consider our competitors. It could be that the experience of these companies with their advertising campaigns may be different from ours.

Despite these concerns about potential sources of error, we are confident that our proposed advertising budget will lead to a successful advertising campaign. We discussed our findings with our advertising consultants, and they confirmed that most of our findings matched their experience. They were slightly surprised by our fitted effect of market capitalisation, but it became clear in the ensuing discussion that even in the advertising community, there are very different opinions about it's effect, and there is no consensus. Therefore, whenever there is a consensus about the qualitative effect of a predictor, our fitted model was in line with that consensus.

write an executive summary for this report.

The purpose of this report is to determine how best to allocate our $2.2 million advertising budget in order to maximise our sales. We gathered a dataset of 490 campaigns run by competitor companies (after excluding several campaigns with particularly unreliable data). For each campaign we estimated the total spending on six different types of advertising: targeted online; general online; targeted television; general television; targeted other; and general other. We also collected data on the company (market capitalisation, headquarter location, years in business, reputation and line of business), the advertised product (quality, style, weight and customisation), and the type of campaign (general, new product or

promotion). We assessed the success of each campaign by comparing the company's sales of the product for the periods preceding and following the campaign. It was not possible to get the exact data on other companies' spending, so we used the estimates of our advertising consultants.

Based on our analysis, we estimate the optimal allocation for our upcoming campaign is $1.2 million on targeted online advertising; $700,000 on general television advertising; $200,000 on targeted television advertising; and $100,000 on general other advertising. We reached these numbers by modelling the relative sales increase as a product of separate functions of each predictor, and another function of the product of product quality and company reputation. To confirm that this model is sufficiently accurate, we compared the predictions with a more flexible model called Random Forest, and found the difference was not significant.

The qualitative findings of our model were also consistent with experience, identifying targeted online spending, type of campaign and product quality as the most important predictors of the success of a campaign. The types of relations identified between the predictors and campaign success were also consistent with experience.

We have tested the robustness of our proposed budget allocation to a range of violations of our assumptions, and found that the proposed budget is expected to be successful across a range of cases. We have identified three main risks to the success of our proposed budget: systematic underestimation, in our modelling, of click-rates for targeted online advertising campaigns; changing trends in adversising, meaning that campaigns that would have been successful in the past no longer attract customers; and sampling bias in our choice of competitor companies.

2. *The following quotes come from a report on predicting the effect of temperature on algal blooms. Where in the report should they be placed? Justify your answers.*

(i)

Algal blooms cause an estimated total annual loss of $1.2 billion due to containment and treatment expenses, costs associated with finding alternative water sources, loss to fisheries, livestock and various other sources of loss. [Green & Pond,2018]

This is probably in the introduction. It is providing context for the problem, motivating the importance of the study. It could also be from the executive summary or abstract, as it is a key part of the context for the problem.

(ii)

There are a few years where average summer temperature was calculated over the period April–July, rather than May–August, due to changes in the definitions in the government environment office. We have excluded these years.

This is from the Data Exploration section. This is clearly part of cleaning the data to ensure it is reliable. This must be done prior to data analysis.

(iii)

We fitted a random forest model with 500 trees. We used 10-fold cross-validation to select the value of the `mtry` *parameter, using the smallest value for which the cross-validated MSE is within one standard deviation of the lowest cross-validated MSE.*

This is from the data analysis section. It is describing the methods used to analyse the data. It is providing all the details needed to repeat the analysis. A brief summary of the methods used to analyse the data might be included in the executive summary or conclusions section, but not with such a level of detail.

(iv)

Because of the correlation between temperature and rainfall, there is a negative correlation between temperature and algal blooms. However, when we correct for the effect of rainfall, the GAM shows that increased average temperature actually decreases the frequency and intensity of algal blooms.

This is from the Conclusion section. It is summarising the results and describing the most important interpretation of these results.

(v)

To determine whether the changes to measuring rainfall in 2009 could have led to incorrect conclusions, we performed a simulation study. We added a standard normal random variable to all rainfall measurements in 2009 and reran the analyses to see whether the results were significantly different. We performed this simulation 100 times. We found that this change increased the estimated relation between temperature and algal blooms by 12%.

This would eiter be in the data analysis section or in the appendix. It describes an analysis performed to check additional details about

the analysis and data issues. Often, this sort of analysis does not need to be included in the main report, and can be included in the appendix. If the potential issue identified is significant then it might be included in the main report in the "Data Analysis" or "Results" section.

(vi)

There is clear evidence that increased temperature with other predictors kept constant is reducing the frequency of algal blooms, but increasing the severity. Because increased temperature often also causes increased rainfall, the overall effect of climate change on algal blooms is difficult to predict.

This is probably from the "Abstract" or "Executive Summary". It is very briefly describing the main conclusions of the report. It might also come from the "Conclusions" section, since it relates to the conclusions of the data analysis.

(vii)

Previous studies relating temperature and algal blooms were based on incomplete meteorological data which did not include rainfall or NO_2 levels (for example [Bloom, 2016] and [Lake, 2018]). By studying the more complete dataset in this analysis, we hope to better adjust for the confounding effects of these predictors.

This is probably from the introduction. It discusses previous work, and is setting the basic context for the data analysis. It is too detailed for the abstract or executive summary.

(viii)

Figure 7 shows the forecast frequency and severity of algal blooms using the fitted random forest model and the climate models from [Sun, 2021]. We see that under the optimistic predictions, the rate of algal blooms is expected to approximately double by 2080. Under the more pessimistic predictions, the rate of algal blooms could double by 2065.

This is from the "Data Analysis" section. It describes a detailed result of a modelling method, and the consequences. It could be in the "Conclusion" or "Discussion" section, but it is unusually detailled for that section, which would usually not include figures.

3. *A data scientist has analysed the data in the file* `HW5Q3.txt` *using the commands in* `HW5Q3_analysis.R`*. The data show the effect of changes of CEO on share prices.*

Variable	*Meaning*
industry	*The industry of the company*
outgoing.ceo.term	*The length of time the CEO was at the company*
outgoing.ceo.gender	*The gender of the outgoing CEO.*
incoming.ceo.gender	*The gender of the incoming CEO.*
date	*The date at which the CEO changed*
month.1.price	*The percentage change in share price in the month following the change.*
month.1.index	*The percentage change in the corresponding market index in the month following the change.*

She has concluded the following:

(a) The 1-month return decreases as the outgoing CEO term increases.

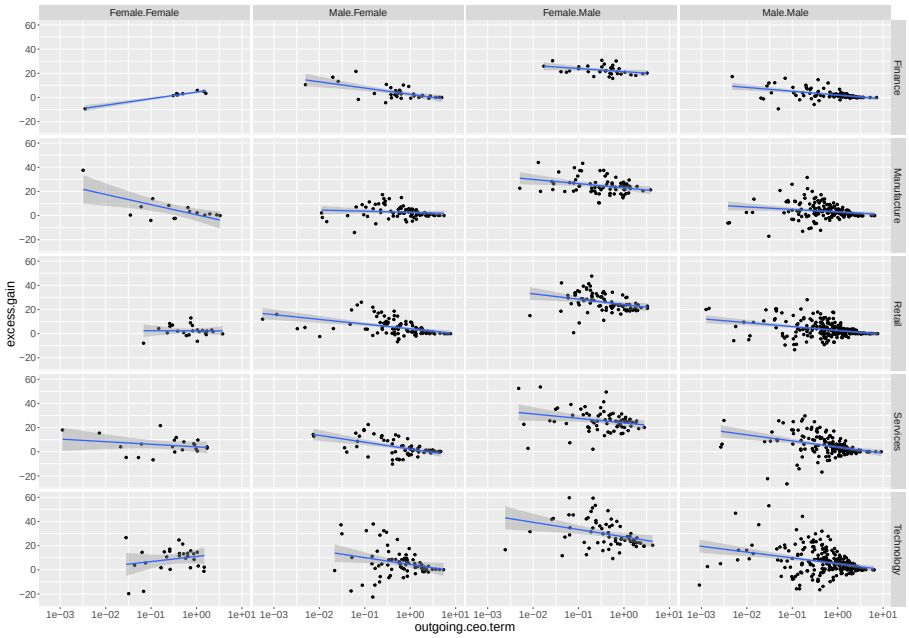
(b) The variance of the 1-month return also decreases as the outgoing CEO term increases.

(c) The 1-month returns are better when the CEO changes from female to male.

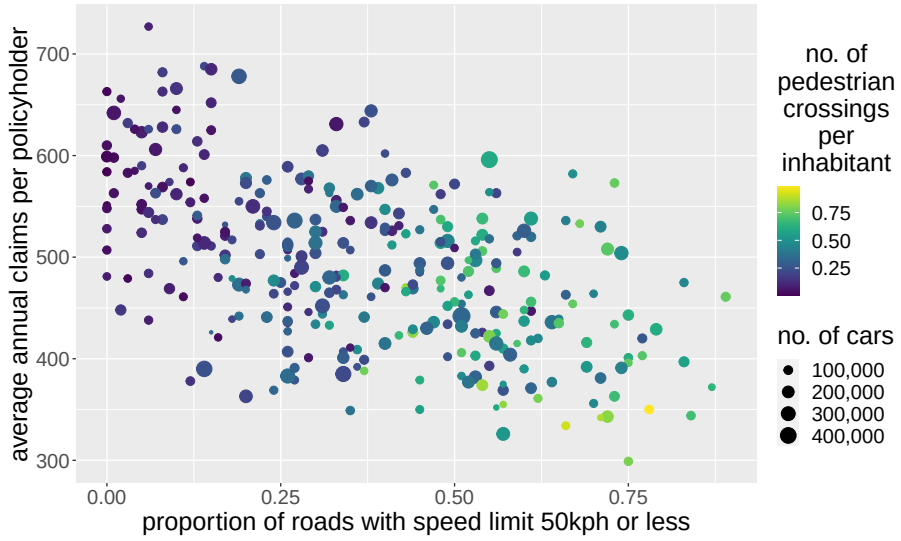
(d) The importance of the outgoing CEO term is different between different industries, being most significant for technology companies, and least significant for Finance companies.

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

Because of the large correlation between the index and the company return, it makes sense to measure the difference between the company return and the index return, and plot this against the predictors. [It probably makes sense to think of the returns multiplicatively, so the return relative to the index would be (1+month.1.price)/(1+month.1.index), but the difference should be close enough to show the pattern.] Using a `facet_grid` for industry and the interaction between incoming and outgoing ceo gender shows these patterns well. We can use the x coordinate for `outgoing.ceo.term`. This probably benefits from a log transformation. We can also add a trend line to each facet. This shows the pattern fairly well.



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 an auto insurance company, and the purpose of the study is to determine the association between safety regulations and accidents.*



Variable	*Meaning*
no.cars	*The number of cars in the city*
population	*The population of the city*
pedestrian.crossings	*The per-capita number of pedestrian crossings*
limit.50	*The proportion of city roads with a limit 50kph or less*
hway.limit	*The speed limit on highways (kph)*
pedestrian.zones	*The proportion of the city centre that is limited to pedestrians*
annual.ave.claims	*The average annual claims for each driver*

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

From Figure 1, we see that proportion of roads with speed limit 50 kph or less and per.capita no. of pedestrian crossings are both negatively associated with annual average claims. Although these two predictors are strongly correlated, we can see that if we fix one, the other still has an impact on average claims. The relation between number of cars and the other variables is less clear.