**Test: Marketing Team**

As a marketing analyst for Pepsi-Cola, you are asked to build an understanding of the sales return of the brand’s marketing mix decisions. To that end, information on weekly brand sales of Pepsi and Coca-Cola are collected over a period of 110 weeks.

The data cover real-world sales and marketing mix information on Coca-Cola and Pepsi-Cola aggregated across a selection of grocery stores at 1 European grocery retailer. The Excel file “MA\_assignment\_data.xls” contains the data on both brands. Brand 1 in the data refers to Coca-Cola, and brand 2 refers to Pepsi-Cola. For example, “sales.brand1” (column B) refers to the sales of Coca-Cola, whereas “sales.brand2” refers to the sales of Pepsi-Cola (column K).

The objective of the analysis is to build an understanding of Pepsi’s sales drivers (not Coca-Cola’s drivers). In other words, your response variable will be “sales.brand2” (not “sales.brand1”).

**Brand 1: Coca-Cola**

**Brand 2: Pepsi-Cola**

The following variables are available for each brand:

A screenshot of a computer

AI-generated content may be incorrect.

**Questions to answer from the data**

1. Calculate descriptive (summary) statistics about the sales and use of marketing mix variables for both brands. Who is the market leader? How do the brands compare in terms of pricing, promotion, assortment? How do the brands compare in terms of allocation of spending between the four advertising instruments? Report the statistics and discuss your learnings and insights.
2. Estimate the following log-log regression model, which explains Pepsi’s sales (“sales.brand2”) as a function of Pepsi’s own marketing mix variables. Note that a value of 1 is only added to variables that contain zero values (feature, display, and the 4 advertising instruments).

(a) log(sales.brand2) = β0 + β1 log(feature.brand2+1) + β2 log(display.brand2+1) + β3 log(assortment.brand2) + β4 log(price.brand2) + β5 log(tv.brand2+1) + β6 log(digital.brand2+1) + β7 log(ooh.brand2+1) + β8 log(magazine.brand2+1)

Report the estimates from equation (a) in a table. Based on this model, what do you conclude about the relation between Pepsi’s marketing mix variables and its sales? How do you interpret the size of statistically significant estimates from equation (a)?

1. The brand would like to understand whether its TV, OOH, magazine, and digital advertising have a longer-term effect on its sales. Estimate the same model as estimated in question 2, but now use an adstock specification for all four advertising instruments, setting lambda to 0.6 for all four advertising instruments. Report the results of the model, interpret the estimates about the impact of advertising on sales and about the impact of all other marketing mix elements, and discuss your findings.
2. Are there any potentially important variables missing from the model? If yes, which ones? What are the implications of these excluded variables for the validity of the estimates presented in Question 3?
3. Assess the predictive ability of the model presented in Question 3. To what extent can the model accurately predict future outcomes?
4. Imagine there have been discussions within the marketing department on whether or not to reduce the share of the total advertising budget allocated to TV advertising with the objective to increase sales. Some members of the department argue that spending should be shifted away from TV to digital advertising. Other members are sceptical about decreasing spending on TV advertising, arguing that TV advertising remains important because it drives sales and reduces price sensitivity. Discuss your view on the matter based on the available data. Do you agree? Why (not)?
5. Management would like to use your model to make a prediction about future sales to better manage inventory levels. You are asked to analyse whether a (single) decision tree would provide a better fit compared to the log-log model. Use the 8 marketing mix variables (without log-transformation) of Pepsi (as in Q3) as input variables, and Pepsi sales (without log-transformation) as the outcome variable. Estimate the model on an 80% training set and 20% holdout set. Use the adstock specification for advertising (same value of lambda as in Q3). Interpret and assess the predictive ability of the decision tree and discuss whether and why this may higher/lower than what you obtained in the log-log model.
6. Estimate a random forest (based on 100 trees), using the same input variables and output variable as in Question 7. What are the most and least important variables and how does this compare to insights obtained by the log-log model in Question 3? Discuss how the fit of the random forest compares to the fit of the decision tree in Question 7.

Format

* R Code used for each question
* Summary of answers per question