MACHINE LEARNING AND OPTIMISATION

Boats: a Segmentation Case

MIM 22 Group 23 – E2

Alexandra Magdei, Jalel Mohib, Jerry Mao,

Margherita Mayr and Rijul Sharma

Structuring the problem

What is the problem that Mary is seeking to address?

- Mary is Senior Manager of Customer Insights at CreeqBoat.
- Because of a crisis in the boating industry, she is analysing different growth strategies for the firm.
- She therefore has to address a **growth problem,** specifically, how to enter the North American market and build more targeted boats for the key customer segments there
- She has to **segment the market** of CreeqBoat's current and potential customers and **identify the key purchase drivers** for boats.

Why is she using PCA, clustering and classification?

- We will be using supervised learning to **predict which customers will buy the boat** (label).
- The dataset contains too many features and some of them might have limited significance when it comes to explaining the label. Therefore, it is important to **first use PCA** to create a dataset with fewer features while minimizing information loss.
- Once PCA has been performed, Mary will **use clustering to segment the database** of its customers and identify which ones CreeqBoat should focus on.
- Lastly, after identifying the most important segments, Mary will use **classification to predict** which customers will buy the boat.

Which questions is she using as input data for each method?

- **Principal Component Analysis**: Q1 has too many features, some of which might not contain important information. Using PCA will create a dataset with fewer features while minimizing loss of information.
- **Clustering:** Q2-Q15 contain demographics information, which will help segment our customer base and get a better understanding of their common characteristics.
- **Classification:** Q16 contains features and Q17 & Q18 contain labels that will be used when building the classification tree.

Sub-questions in Q1

Are the answers to the questions correlated? Do the correlated questions make sense if you look at the questions in Appendix 1?

Answers are not highly correlated, no observation has a correlation higher than 0.7 in the dataset.
If we use 0.5 as a correlation threshold, we have Q1.26 correlated with Q29. Looking at Appendix 1, this does make sense because they are about conspicuous consumption.

Do we need to scale and/or normalize the data?

We don't need to scale the data because the questionnaire is already scaled on a scale of 1 to 5.

How could this be useful in a marketing survey?

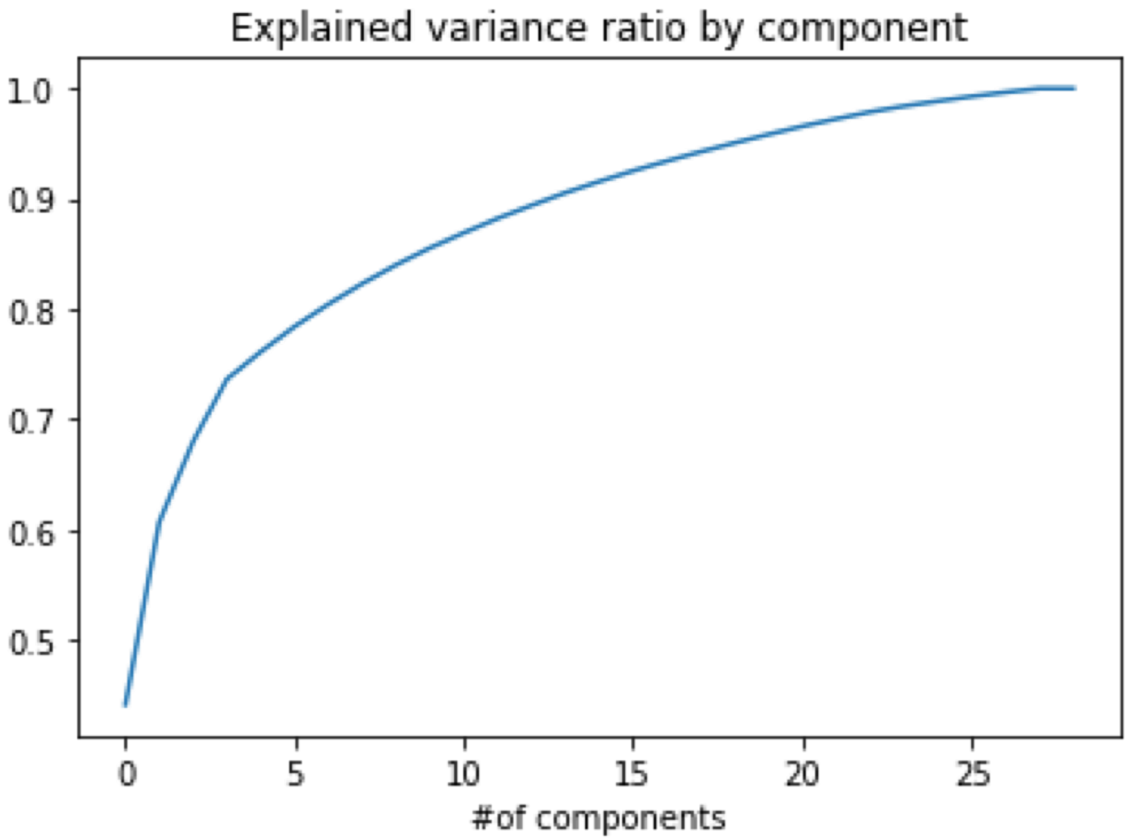
The way the questions have been phrased might make respondents answer in a particular way, and therefore reiterating the attitude/beliefs that we are trying to uncover by rephrasing them will help us avoid these biases.

Number of features selected

In order to choose the optimal number of features, we need to plot the **explained variance ratio curve** and understand how the number of components varies with the explained variance in components.

We noticed that **selecting 5 features explains ~79% of the variance** in the dataset. From 10 components onwards, the increase in explained variance is not significant.

Fitting a PCA model to the data



SparsePCA

Meaning of each new component

Component 1: Q16, Q17, Q20, Q23, Q27, Q28
The negative correlation suggests that people are not knowledgeable of technical information regarding boats.

Component 2: Q2 & Q12
Price sensitivity is high among respondents.

Component 3: Q1, Q6, Q7, Q18, Q19, Q21, Q22, Q24, Q25
Boats have an emotional and social value.

Component 4: Q11
Respondents prefer to do maintenance & repair on their own (implication: we won’t be able to sell them insurance services).

Component 5: Q3, Q5, Q9, Q13, Q14, Q26
Quality and brand are important.