450 Solo 1 Report

R Sangole

2018-07-16

Introduction

This paper is organized as follows.

Secion I deliniates the overview of the methodologies used and the challenges faced at a high level. It also explains some technical challenges faced.

Section II explains the data preparation activities.

Section III outlines details of t-SNE approach.

Section V outlines details of k-Means approach.

Section VI outlines details of the PAM approach.

Section VII outlines details of the Model-based clustering approach.

Section VIII talks about the Market Segmentation profiling.

Section IX outlines how to perform classification on the model.

Section I - Overview of Methodologies Used

Section II - Data Preparation

The original dataset available for analysis comprises of responses from 1800 customers for 16 questions. These questions span objective multi-choice demographic questions (age, gender, education, income) to personality and personal preference related questions on a Likert scale. Since the task at hand is to develop an *attitudinal post hoc segmentation*, it's important to first cleanse these data to responses which are relevant in this analysis. Furthermore, it's important to quality check these data against some rules while also addressing cases of missing values. This section describes the modifications made on the original data.

Data Modifications

RULES There are some inconsistencies in the data which were corrected by simple rules.

- Rule A - If q4 r11 is true, it indicates that the respondant doesn't use any apps. If this is the case, then q11 should be None, and q12 should be blank.
- Rule B - To preserve ordinality of q11, 'none' is set to 0, instead of 6
- Rule C - For responses in q11 where the respondant says 'Dont know how many apps', I've set these to NA, so they can be imputed later.

- Rule D - For q12 (% of free apps), there are values missing (when q11 is None), which are set to 6 (All free apps). This will allow these rows to be used in the clustering.

MISSING VALUES Once the rules are applied, There are 99 missing values in q57 and 53 missing values in q11. Imputation is carried out using the `mice` package using a random forest method.

RECODING A significant amount of recoding was done on most of the questions. For example:

- Q13 - Website visit frequency:
  - Social Visit Freq = Average of Facebook, Twitter, LinkedIn and Myspace
  - Music Visit Freq = Average of Pandora, Vevo, AOL Radio, Last.fm and Yahoo music
  - Video Visit Freq = Average of Vevo, YouTube, and IMDB
- Q24 - Technological Sentiments: The 12 questions are summarized into a few attitudinal basis variables:
  - Positive attitude towards technology
  - Entertainment as a primary use of technology
  - Communication as a primary use of technology
  - Negative view of technology
- Q25 - Personality related questions are grouped into 4 main themes:
  - Leadership view of self
  - Risk taker personality
  - High drive towards life
  - Follower
- Q26 - Shopping trend related questions are grouped into five themes:
  - How important are bargains?
  - How important are brands?
  - Do you believe one earns money to spend on oneself?
  - How much do you love apps?
  - Do children influence your purchases?
- Q2 - Platforms - Apple, Andriod, Windows, or Other

Almost every original question is modified to more usable and succint groupings.

SUBSETTING Iteratively, a total of 27 key variables were taken into the analysis going forward.

Section III - t-SNE

t-SNE, or t-Distributed Stochastic Neighbor Embedding is a non-parametric technique to perform dimensionality reduction suited for high-dimensional large datasets. It maintains the underlying structure (local variation) in higher dimensional data while also capturing the macro-structure of the data. t-SNE has been used for visualization in a wide range of applications, including computer security research, music analysis, cancer research, bioinformatics, and biomedical signal processing. It is often used to visualize high-level representations learned by an artificial neural network. Upon application of t-SNE to the full multivariate dataset, we obtain a 2-dimensional representation as shown to the right.

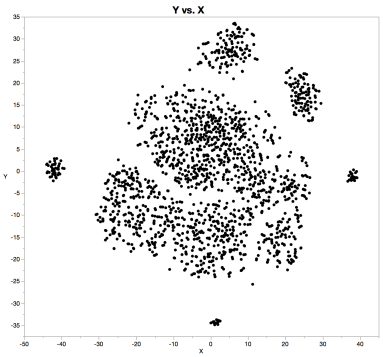

Figure 1: t-SNE representation

This plot can be overlayed with, or colored with the any of the explanatory variables in our dataset to gain insights into the structure of these data. For example, if the plot is overlayed with the variable for race, we can see some remarkably clear distinctions in the data:

- A, B, C = White
- D = African American
- E, G = Latino
- F = Asian
- H = Hawaiian

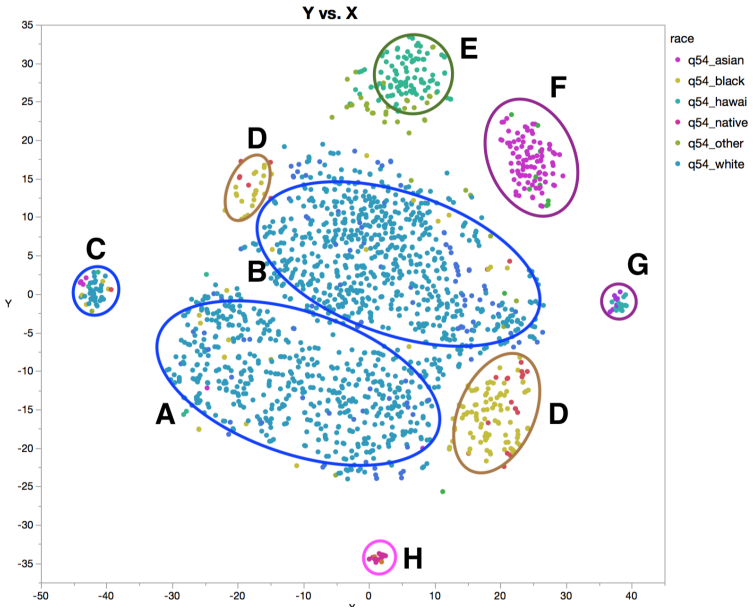

Figure 2: t-SNE with race

There are many more insights which can be quickly sought through analyzing multivariate data through t-SNE. While it's tough to represent all of them visually in a report, this hand representation attempts to explain the clusters observed:

Some of the key takeaways:

- ~ 100% of Asian respondants use Apple devices
- ~ 0% of Black respondants use Windows phones, and ~0% of Black respondants use tablets
- Cluster E - 100% latino cluster is largely an Apple device user, with ~0%windows usage
- Cluster C - 100% of this cluster do not use any apps
- Cluster C is also mostly White, 60 years +, and richer
- A majority of Android users are White
- Everybody plays games regularly, regardless of gender, race, or age
- Younger crowds are more brand aware than older crowds
- A large majority of Asian users are TV related App users, and music related app users
- Irrespective of gender or marital status - brand awareness, app lovers and belief in earning money to spend on oneself go hand in hand
- Most folk think of themselves are thought leaders
- As age increases, folks tend to be more risk averse

Section VII - Wrap Up

References

1. https://lvdmaaten.github.io/tsne/
2.