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 some of the exploratory work done.

Section III explains the data preparation activities.

Section IV 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 - EDA

UNIVARIATE STUDIES Time series plots were run for all the variables to get an idea of the underlying structure. While some signals don't show strong seasonal patterns like in figure 1. Others show very strong seasonality, like in figure 2. Depending on the chosen solution, this is useful information. The response variable `total_cases` shows the peaks and available information for the two cities. Note teh different time scales on the x-axis.

Section III - Data Preparation

Section IV -

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.

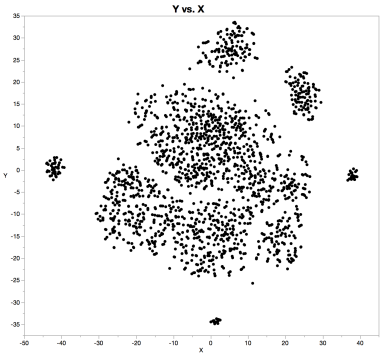

Figure 1: t-SNE representation

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

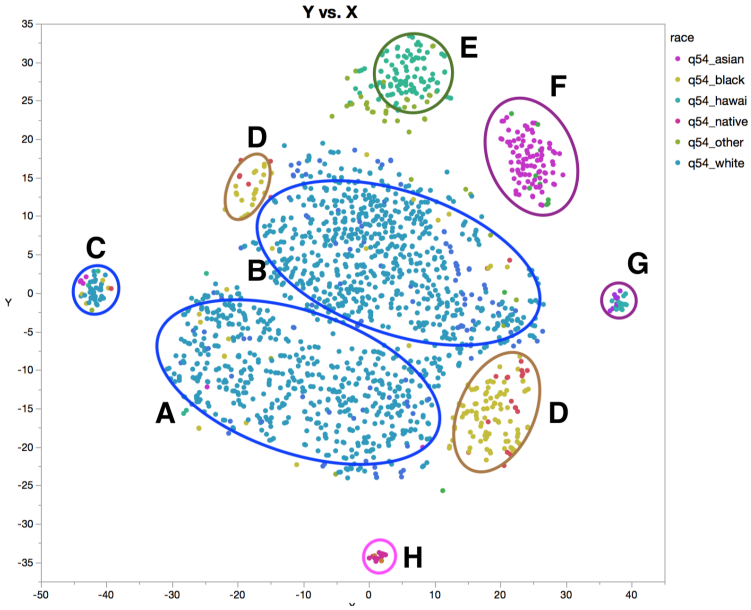

Figure 2: t-SNE with race

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.