Dentsu Aegis Data Event

Ryerson Applied Math M.Sc Group

January 30, 2020

Contents

Chapter 1

Preliminary Thoughts

1.1 Introduction

I set this up for us to have a place to write any notes before going into the event and during so we can share any helpful tool and resources we have. From the little reading I've done it seems like we'll be looking at an online marketing exercise. Most of the services they've provide have been using AI and ML algorithms in order to provide targeted ads (eg. using face recognition on social media to determine which people have pets and push pet food ads through their feed).

- Alin

1.2 Logistic Regression

A very well known quantity in online marketing campaigns is logistic regression. It uses the logistic function in order to determine the likelihood of success (ad being appropriate for a given person) and then based on a threshold makes a binary decision on whether or not you should push the ad out to the user. This can be extended to a multinomial logistic regression where we can compute likelihoods for multiple categories (use threshold to say yes, maybe, no) on the ad and then perform a constraint optimization factoring in costs of pushing ads.

The model is set up as follows.

- Let $p = \mathbb{P}(X = 1)$ be the probability of a success
- Let $\{x_1, \ldots, x_n\}$ be the covariates

- Let $\beta_0, \beta_1, \ldots, \beta_n$ be the coefficients
- note that logit : $\mathbb{R} \to (0, 1)$ is defined as $\text{logit}(x) = \frac{1}{1-e^{-x}}$

With the regression equation being

$$\text{logit}(p) = \beta_o + \beta_1 x_1 + \ldots \beta_n x_n \tag{1.1}$$

Fitting of this model can be done in several ways with the most common being the maximum likelihood estimator (MLE). An example of logistic regression in R via the `glm` function along with several tests for assessing the quality of the fit.

1.3 Possible "Solutions"

Below are two standard applications of ML in digital marketing. Several algorithms can be used to solve these problems which will be detailed as needed but in the simplest case can be solved with the logistic regression model above.

1.3.1 Content Curation

This problem is one of filtering. Provided with some information about a user, we should be able to eliminate "annoying" content/ads from their experience while at the same time providing ads that are most likely to drive business.

1.3.2 Product Recommendation

In simple terms, this is a matching exercise. Provided with some information about a client, we should be able to offer some products that they are more likely to be interested in. Solving this problem can be done using a multinomial logistic regression where the dimension of p (the vector containing probabilities of "success") is equal to the number of items in your inventory. The model should be trained on previous purchasing data. It is worth noting that majority of the covariates will be factor variables so regression may appear to be a poor fit.

Chapter 2

Data Work

2.1 Background

We have a cannabis brand looking at the effects of legalization on cannabis consumption and the consumer attitude within that. The goal is to create a narrative around this idea that can be given to stakeholders.
Contact: email by 1:00pm to confirm participation

2.1.1 Initial Reaction

At first glimpse, this project seems like it should be approached as a *sentiment analysis*. We should try to do some text mining in order to augment the data given already. Here is an example of a sentiment analysis in R using tidy data (this is a clear definition which will guide the data cleaning process).

2.2 The Dataset

THe data comes in several files looking at different aspects of the cannabis market. These will be combined in order to give insight into how the market feels about the product and potential habbits. All of the data is in the form of a cross-tabulated table providing the summary of a survery adminitered in 2018 and 2019 respectively (with the 2019 being more compreshensive).

2.3 EDA

2.4 Analysis of Data

2.4.1 Annual Changes

For this section we track the movements of people's votes by using the 2018 survey as a baseline. This can be done via a straight differencing.

2.4.2 PCA on Various directions

Done by Kenji

2.4.3 Survey Assessment

Since the data collected includes questions, these themselves should be analyzed in order to get an idea of any bias that was introducted into the data via influential language. The first column of the data set can be used in order to do a sentiment analysis (a scoring system is used to see if we have overly negative or overly positive language). The outcome of each question, and the survey as a whole can be recorded. A thresholding system can be implemented on the differences to see if perception really has improved over the years (this should be particularly interesting in the proportions of "older" people). The idea here is that if we have a highly negative question, even small improvements can be seen as impactful (maybe more impactful overall) and thus should be reweighted.

This can all lead into the creation of a metric tracking the perception of the cannabis industry based on the results of this survey year to year as well as any additinal revenue data we can get our hands on. Mathematically, this will be a weighted average and displayed in a nice big font to make it look like we "scored" the market.

2.5 Insight and Solutions

To be done in R Markdown or Jupyter Notebooks. Here is a an extra step.