State University of New York at New Paltz

Aniruddha Tidke

DATA ANALYTICS PROJECT

CPS593-02 “Data Analytics”

(Professor Dr. Chirakkal Easwaran)

Summer 2020

TABLE OF CONTENTS

QUESTION:

Consider the height and weight data given in height_weight.csv file (This is same data you have seen before but there is no size preference column). Suppose you are in a t-shirt manufacturer, deciding how many sizes of t-shirts (S,M,L,XL) to make (Not the numbers in each size, but how many sizes – should you make S, M only, or S, M, L only

Or all of S, M, L, XL or some other combination). The goal is to ensure good-fitting t-shirts, at the same time maximizing profit.So the question for you is to decide: How many different t-shirt sizes should you produce for the customers in the dataset ?

This is an open ended question. There is no right or wrong answer. You can make reasonable assumptions about t-shirts such as

1) Generally larger people prefer larger t-shirts, and that

2) Bigger sizes sell at a slightly higher price than smaller sizes.

And potentially other reasonable assumptions.

Based on your assumptions, argue why making a certain combination of sizes make most sense, given your goal. Again, this is not a question with right or wrong answer.

Answer:

What t-shirt size would you send to someone if you don’t know their shirt size, but instead you know their height, weight, and gender?

Electronic shopping gives unfathomable solace to customers, regardless, it furthermore brings issues, for instance, size fitting due to no rapidly trying it out. Without trying it out, picking the size of the thing is one of the issues that impact customers' shopping experience, which in like manner prompts high return rates to associations. Things' size, health desire and size proposal are along these lines essential for web shopping. In this paper, we pondered the size desire issue and proposed a strategy to improve the prediction accuracy. We get the unnoticeable semantics of customer reviews by using Bert model to get comfortable with the inert depictions. Significant features are discovered from the overview data, solidifying with customers' body features and the thing's information to predict the thing's availability for a customer. The tests on the certified web attire retailer dataset endorse the reasonability of the proposed procedure.

First I would need to show the issue and perceive data to manufacture the model. What are the data sources and the yields? How do the wellsprings of information map onto the yields?

I understood that the primary information I'd have to envision shirt sizes from was: sexual direction, height and weight. Once in a while I'd have a weight list (BMI) as opposed to height and weight.

I in like manner understood that I'd send "unisex" shirts. I've done a great deal of online shirt shopping, so it became clear to me to look at how electronic clothing stores deal with this issue – enter the "size blueprint". Okay, so now I'd found an estimation that could be used to arrange shirt sizes: chest size.

So now I essentially expected to make sense of how to get from sexual direction, height, and weight to chest size.

I know in actuality that there are a couple of immense open datasets that contain prosperity and prosperity related estimations.

Clearly, they would have sexual direction, height, weight and BMI. It would be an altogether odd prosperity and sustenance concentrate in case they didn't have fundamental economics and estimations. In any case, chest estimation? The setting had all the earmarks of being to some degree odd. What does chest limit have to do with prosperity and sustenance?

Maybe chest edge is huge in the event that we're taking a gander at unwinding? Aha! It has chest estimations for internal breath and exhalation as a significant part of the joint irritation datasets!

The last piece of the puzzle was recognizing a lone year accomplice that contained these measures together, in light of the fact that I would require the data to be at the level of unique individuals. I found these estimations were contained in the 2009-2010 friend, so I downloaded the dataset records for that accomplice.

These are the variables I'll pull out from the different data frames.

- SEQN - Participant ID
- ARXCCIN - Inhale chest circumference in CM
- BMXWT - Weight in KG
- BMXHT - Height in CM
- BMXBMI - Body Mass Index (BMI)
- DMDHRGND - Gender of participant (1 = Male, 2 = Female)

I'm using the inhale chest circumference because I found in some exploratory analyses (not reported here) that it has a stronger correlation to the other measurements. I guess the exhale circumference was more noisy for some reason?

Online shopping is getting standard at a confusing pace for its advantage of shopping wherever we like without setting off to the hypostatic store. In any case, standing out from disengaged shopping which engages people to pick sensible pieces of clothing by trying, online shopping just allows customers to pick articles of clothing by examining pictures, depictions and seeing size charts. Beside the nonattendance of assess choices, differing size standards across different brands experience furthermore become a difficulty for customers while picking the right size. When in doubt, different brands of dress may have assorted size rundown arranging standards, for example, the M size T-shirt of brand A, may have a comparable length as the L size T-shirt of brand B. Likewise, regardless, dress of a comparable brand have unmistakable size standards, for they may start from various item contributions. These real factors lead to a vulnerable shopping experience when people every now and again pick a wrong size.

Model Building:

I decided to manufacture two models in light of the fact that sometimes I know a person's height and weight anyway not their BMI, and various events I know their BMI yet not their height and weight.

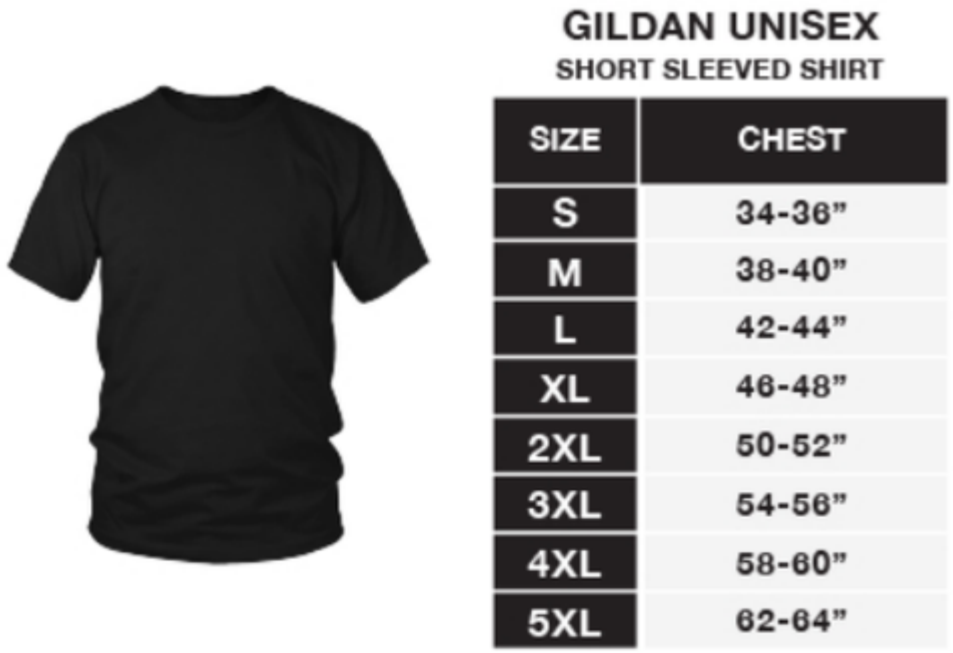
Regardless of the way that I could take the total of the heights and loads and convert them into BMIs, I'm theorizing this could lead to lost information. The association above recommends that weight is the best pointer of chest size (better than BMI), so uncovers to me it might be more brilliant to use height and weight when it's available, anyway use BMI as a fallback when stature and weight are not open.

Height and weight model:

To gather a model, I'll run a stepwise immediate backslide to choose the model of best fit. I'll enter height, weight, and sexual direction as markers and license correspondence terms. I'll use AIC (Akaike's Information Criterion) for model decision, since it rebuffs models with included multifaceted nature and I need a stingy model that doesn't overfit the data.

I see that the best model consolidates all terms, including weight*height and weight*gender association terms.

Chest size to shirt size:



GILDAN UNISEX
SHORT SLEEVED SHIRT

SIZE	CHEST
S	34-36"
M	38-40"
L	42-44"
XL	46-48"
2XL	50-52"
3XL	54-56"
4XL	58-60"
5XL	62-64"

For no good reason the chart leaves "XS" and the degrees lso don't give full incorporation of the possible chest size characteristics (34-36" and thereafter 38-40"???). So I'll expand the upper bound of each range to meet the lower bound of each size above it. This will cause the gauges to fizzle on the size of greater sizes rather than smaller sizes, which I accept is better assuming that a shirt is excessively huge, in any occasion you can regardless wear it!

R-Language Code For T-Shirt Size Prediction

```
# First we will set up the directory
#setwd()

library(SASxport)
library(tidyverse)
library(MASS)

a <- read.csv("height_weight.csv")

a %>% dplyr::select(id = SEQN, chest_in_cm = ARXCCIN) -> chest_measures
b %>% dplyr::select(id = SEQN, weight_kg = BMXWT, height_cm = BMXHT,
bmi = BMXBMI) -> height_weight
d %>% dplyr::select(id = SEQN, gender = DMDHRGND) %>%
  mutate(gender = case_when(gender == 1 ~ "M", TRUE ~ "F")) -> gender

# We will join datasets and select only the rows that have all
measurements
chest_measures %>%
  left_join(., height_weight, by = c('id')) %>%
  left_join(., gender, by = c('id')) %>%
  filter(!is.na(chest_in_cm),
         !is.na(height_cm),
         !is.na(weight_kg),
         !is.na(gender),
         !is.na(bmi)) -> df


## Correlations
cor(df$weight_kg, df$chest_in_cm)
cor(df$height_cm, df$chest_in_cm)
cor(df$bmi, df$chest_in_cm)


df %>%
  group_by(gender) %>%
  summarize(mean = mean(chest_in_cm))


## Model Building
### height and weight model

lm(data = subset(df, select= c(chest_in_cm, height_cm, weight_kg,
gender)),
   chest_in_cm ~ .) -> mod
```

```
step.model <- stepAIC(mod, direction = "both", trace = TRUE, scope = .
~ .^2)


summary(step.model)

# Save the model for later
height_weight_model <- step.model

lm(data = subset(df, select= c(chest_in_cm, bmi, gender)), chest_in_cm
~ .) -> mod
step.model <- stepAIC(mod, direction = "both", trace = TRUE, scope = .
~ .^2)

summary(step.model)

# Save the model for later
bmi_model <- step.model

mean(df$weight_kg)
mean(df$height_cm)
mean(df$bmi)



input <- data.frame(height_cm = 168, weight_kg = 83, gender = "F")

# 1 cm = 0.393701 inches
predict(height_weight_model, input) * 0.393701


input <- data.frame(bmi = 29, gender = "F")
predict(bmi_model, input) * 0.393701


input <- data.frame(height_cm = 168, weight_kg = 83, gender = "F")

data.frame(chest = predict(height_weight_model,input)[[1]] * 0.393701)
%>%
  mutate(shirt_size = case_when(
    chest < 32 ~ "XS",
    between(chest, 32, 36) ~ "S",
    between(chest, 36, 40) ~ "M",
    between(chest, 40, 44) ~ "L",
    between(chest, 44, 48) ~ "XL",
    between(chest, 48, 52) ~ "2XL",
    between(chest, 52, 56) ~ "3XL",
    between(chest, 56, 64) ~ "4XL",
    chest > 64 ~ "5XL"
  )
  )
```

```r
## Wrapping all in a function

predict_shirt_size <- function(height_cm, weight_kg, bmi, gender){

  if(!is.na(height_cm) & !is.na(weight_kg)){
    input <- data.frame(height_cm = height_cm, weight_kg = weight_kg, 
gender = gender)
    chest <- predict(height_weight_model, input)[[1]] * 0.393701
  } else if(!is.na(bmi)){
    input <- data.frame(bmi = bmi, gender = gender)
    chest <- predict(bmi_model, input)[[1]] * 0.393701
  } else {
    # Just in case height, weight, and BMI are all missing
    chest <- -99 # A value that gets ignored below
  }

  shirt_size = case_when(
    between(chest, 0, 32) ~ "XS",
    between(chest, 32, 36) ~ "S",
    between(chest, 36, 40) ~ "M",
    between(chest, 40, 44) ~ "L",
    between(chest, 44, 48) ~ "XL",
    between(chest, 48, 52) ~ "2XL",
    between(chest, 52, 56) ~ "3XL",
    between(chest, 56, 64) ~ "4XL",
    chest > 64 ~ "5XL",

    # Defaults in case height, weight, BMI are missing
    # but gender is not
    gender == "F" ~ "M",
    gender == "M" ~ "L",

    # Last resort
    TRUE ~ "L"
  )

  return(shirt_size)
}

predict_shirt_size(height_cm = 168, weight_kg = 83, gender = "F")
```