Predicting Self-Rated Mental Health Based on Demographic and Family Traits

James Bao, Alan Chen, Xinyi Zhang, Rose

10/19/2020

Abstract

Predicting mental health in the context of the nuclear family. Does having children have a positive (fulfilled) or negative impact (stressed)? Does being married have a positive or negative impact?

Introduction

Data

The dataset we used in our modeling is the 2017 General Social Survey (Family cycle). The following sections will discuss how the data was collected, what the key features of the dataset are, and what the data looks like.

Data Collection

From February 1, 2017 to November 30, 2017, Statistics Canada gathered data on the Canadian family unit by conducting voluntary telephone interviews. Their target population was all non-institutionalized individuals living in Canada, aged 15 or older. Cross-sectional sampling was conducted in a two-stage design.

The stratified simple random sampling method was used in the first stage. Here, the sampling frame consisted of telephone numbers from the Census grouped as households using data from Statistic Canada's dwelling frame. Strata were formed at the census metropolitan area (CMA) level and at the province level (i.e., large CMAs formed their own strata, smaller CMAs were grouped together, and the non-CMA regions of each province were grouped together), forming a total of 27 non-overlapping strata. Finally, households were sampled randomly from each strata such that the number sampled units from each strata corresponded to the population sizes of each strata. To reiterate, the sampled population for this first stage was the chosen households from each strata.

The stratified simple random sampling method was also used in the second stage. Here, the sampling frame was a list of household members, aged 15 and older, from the households selected in the first stage. Then, one individual was randomly selected from each household, forming the sampled population. Approximately 43,000 individuals were contacted to participate in the survey.

Overall, the surveying method using two-stage simple random stratified sampling is effective in generating a sample that geographically representative of the Canadian population. In addition to estimates about the Canadian population at large, the stratified sampling method also allows estimates to be made about subpopulations (at the province level).

Statistics Canada reported that the non-response rate was 47.6%. To reduce the effects of non-response bias, survey responses were adjusted based on the demographic characteristics of households that were non-responsive (by pulling their information from the 2016 Census). This ensures that the discrepancy between the target population and survey responses resulting from non-response is minimized. Furthermore, for the Family cycle of the GSS, responses were also adjusted for income and household size to make more accurate survey estimates for the variables of interest.

Statistics Canada did not disclose the true cost of conducting the survey but we can make some speculations based on the available information about their field work methodology. Surveying was conducted using Computer Assisted Telephone Interviewing (CATI) wherein interviewers read aloud the computerized questionnaire and immediately record the respondent's answers. Although this allows for a reduction in costs compared to traditional in-person surveying, labor costs still include time spent computerizing the survey, training interviewers, and having interviewers administer the questionnaire. Other labor costs include designing the questionnaire and surveying methodology as well as conducting quality control (data consistency was checked by the CATI system during surveying and unresolved inconsistencies were handled afterwards by support staff). Non-labor costs likely included paying for equipment, phone service, offices, and so forth. Again, although we don't have exact costs, we can conclude that the time and costs associated with conducting the GSS is a clear reason why it is only administered once a year.

Per the report on the 2017 GSS from Statistics Canada, extensive research and testing was conducted when designing the questionnaire. Consequently, a major strength of the questionnaire is that it contains focused questions that comprehensively and extensively capture the subject of interest (the Canadian family). Upon reading through the questionnaire made available by Statistics Canada, the wording of each question is precise and clear, leaving little room for ambiguity. Additionally, another strength of the survey is that a vast majority of questions were objective (dates, events, counts) removing potential response biases that occur with subjective questions. (Not all questions were objective however, in fact the variable of interest we will model in subsequent sections consists of subjective responses.) On the other hand, because of the specificity of the questions, the survey is very long with several dozens sections and several questions per section. Furthermore, as a result of the large scope of the target population, many questions in the survey did not apply to a large majority of respondents (e.g., number of grandchildren, questions about additional marriages, etc.). The data collected is also incomplete because participants were given the option to refuse to answer or answer "I don't know" to each question since participation was voluntary.

Overall, the surveying methodology and distributed questionnaire were carefully designed in the interest of collecting accurate, representative data wherever possible.

Data Characteristics

-key features, strengths, weaknesses -discuss variables: similar that you don't use, construct variables by combining various ones

Data Visualization

-what does the data look like -plot the raw data

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------- tidyv
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data <- read_csv("gss_cleaned.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   caseid = col_double(),
##   age = col_double(),
##   age_first_child = col_double(),
##   age_youngest_child_under_6 = col_double(),
##   total_children = col_double(),
##   age_start_relationship = col_double(),
##   age_at_first_marriage = col_double(),
##   age_at_first_birth = col_double(),
##   distance_between_houses = col_double(),
##   age_youngest_child_returned_work = col_double(),
##   feelings_life = col_double(),
##   hh_size = col_double(),
##   number_total_children_intention = col_double(),
##   number_marriages = col_double(),
##   fin_supp_child_supp = col_double(),
##   fin_supp_child_exp = col_double(),
##   fin_supp_lump = col_double(),
##   fin_supp_other = col_double(),
##   is_male = col_double(),
##   main_activity = col_logical()
##   # ... with 1 more columns
## )

## See spec(...) for full column specifications.
```

```
data <- data[, c("age", "sex", "education", "religion_has_affiliation", "marital_status", "total_childre
```

Model

Bayes' Theorem for Naive Bayes Classifier: $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$

```
# using 12 rows of dummy data until we get the real data

data <- tibble(
  sex = c("M", "F", "M", "M", "F", "F", "F" ,"M", "F", "M", "F", "M"),
  martial_status = c("Married", "Divorced", "Single, never married", "Living common-law", "Widowed", "Se
  age = c("under 18", "18-34", "35-50", "50-70", "over 70", "under 18", "18-34", "35-50", "under 18", "
  education = c("Bachelor's", "High school", "College", "Trade Cert", "University degree above Bachelor
  has_regilious_affliction = c("Y", "Y", "N", "Y", "Y", "N", "Y", "Y", "N", "Y", "Y", "N"),
  has_children = c("Y", "N", "Y", "N", "Y", "N", "Y", "N", "Y", "N", "N", "N"),
  selfrated_mental_health = c("Fair", "Good", "Very Good", "Excellent", "Poor", "Fair", "Good", "Very Go
)
data
```

```
## # A tibble: 12 x 7
##    sex   martial_status age   education has_regilious_a~ has_children
##    <chr> <chr>          <chr> <chr>     <chr>            <chr>
##  1 M     Married        unde~ Bachelor~ Y                Y
##  2 F     Divorced       18-34 High sch~ Y                N
##  3 M     Single, never~ 35-50 College   N                Y
##  4 M     Living common~ 50-70 Trade Ce~ Y                N
##  5 F     Widowed        over~ Universi~ Y                Y
##  6 F     Separated      unde~ Universi~ N                N
##  7 F     Married        18-34 Bachelor~ Y                Y
##  8 M     Divorced       35-50 High sch~ Y                N
##  9 F     Single, never~ unde~ College   N                Y
## 10 M     Married        18-34 Trade Ce~ Y                N
## 11 F     Divorced       35-50 Bachelor~ Y                N
## 12 M     Single, never~ 18-34 High sch~ N                N
## # ... with 1 more variable: selfrated_mental_health <chr>
```

```r
# add 1 after all x-y combinations to avoid zero frequency problem

# count y
poor <- data %>% filter(selfrated_mental_health=="Poor") %>% tally() + 1
fair <- data %>% filter(selfrated_mental_health=="Fair") %>% tally() + 1
good <- data %>% filter(selfrated_mental_health=="Good") %>% tally() + 1
vgood <- data %>% filter(selfrated_mental_health=="Very Good") %>% tally() + 1
excellent <- data %>% filter(selfrated_mental_health=="Excellent") %>% tally() + 1

total_mental <- poor + fair + good + vgood + excellent

# count sex given y
male_poor <- data %>% filter(sex=="M" & selfrated_mental_health=="Poor") %>% tally() + 1
male_fair <- data %>% filter(sex=="M" & selfrated_mental_health=="Fair") %>% tally() + 1
male_good <- data %>% filter(sex=="M" & selfrated_mental_health=="Good") %>% tally() + 1
male_vgood <- data %>% filter(sex=="M" & selfrated_mental_health=="Very Good") %>% tally() + 1
male_excellent <- data %>% filter(sex=="M" & selfrated_mental_health=="Excellent") %>% tally() + 1
female_poor <- data %>% filter(sex=="F" & selfrated_mental_health=="Poor") %>% tally() + 1
female_fair <- data %>% filter(sex=="F" & selfrated_mental_health=="Fair") %>% tally() + 1
female_good <- data %>% filter(sex=="F" & selfrated_mental_health=="Good") %>% tally() + 1
female_vgood <- data %>% filter(sex=="F" & selfrated_mental_health=="Very Good") %>% tally() + 1
female_excellent <- data %>% filter(sex=="F" & selfrated_mental_health=="Excellent") %>% tally() + 1

total_male <- male_poor + male_fair + male_good + male_vgood + male_excellent
total_female <- female_poor + female_fair + female_good + female_vgood + female_excellent

# count has_children given y
has_children_poor <- data %>% filter(has_children=="Y" & selfrated_mental_health=="Poor") %>% tally() +
has_children_fair <- data %>% filter(has_children=="Y" & selfrated_mental_health=="Fair") %>% tally() +
has_children_good <- data %>% filter(has_children=="Y" & selfrated_mental_health=="Good") %>% tally() +
has_children_vgood <- data %>% filter(has_children=="Y" & selfrated_mental_health=="Very Good") %>% tal
has_children_excellent <- data %>% filter(has_children=="Y" & selfrated_mental_health=="Excellent") %>%
no_children_poor <- data %>% filter(has_children=="N" & selfrated_mental_health=="Poor") %>% tally() +
no_children_fair <- data %>% filter(has_children=="N" & selfrated_mental_health=="Fair") %>% tally() +
no_children_good <- data %>% filter(has_children=="N" & selfrated_mental_health=="Good") %>% tally() +
no_children_vgood <- data %>% filter(has_children=="N" & selfrated_mental_health=="Very Good") %>% tally
no_children_excellent <- data %>% filter(has_children=="N" & selfrated_mental_health=="Excellent") %>%
```

```r
total_has_children <- has_children_poor + has_children_fair + has_children_good + has_children_vgood + h
total_no_children <- no_children_poor + no_children_fair + no_children_good + no_children_vgood + no_chi

# repeat for the rest of the independent var



## example: predict mental health state given male, has children
# calculating prob(poor)
class_prior_prob_poor <- poor / total_mental
likelihood_poor <- (male_poor/poor) * (has_children_poor/poor)
predictor_prior_prob_poor <- (total_male/total_mental) * (total_has_children/total_mental)
prob_poor <- likelihood_poor * class_prior_prob_poor / predictor_prior_prob_poor

prob_poor
```

```
##           n
## 1 0.1545455
```

```r
# repeat this process to get probability for fair, good, vgood, excellent
# the classification will be the class with the largest probability

# note that the probability value themselves should not be taken seriously since naive bayes is a bad e
```

Results

Discussion

Weaknesses

Next Steps

Appendix

References

Interview method/survey size: https://www.statcan.gc.ca/eng/survey/household/4501 Detailed information about GSS 2017: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816 Questionnaire: https://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=assembleInstr&lang=en&Item_Id=335815#qb345205