Predicts Individual Life Satiafication by Demographic Characteristics

Heran Zhou, Xinyu Zhong, Yaqi Feng, Yuhan Gu

Octobor 19, 2020

github: https://github.com/Kevinzhou0717/STA304-Problem-Set-2

1 Abstract

Canada has long been recognized as a country with high levels of citizen life satisfaction, and factors related to family are thought to be greatly influential to citizen's feelings on life. This study is based on the 2017 General Social Survey ("GSS"), with the central theme being "families", and aims to use the logistic regression model with demographic characteristics to predict the life satisfaction of individuals. Although there were some biases in life satisfaction as a subjective factor, the results of the model provided evidence for the contribution of gender, income, education, family relationship, and other factors to life satisfaction. Government bodies and interested individuals can use the result to evaluate the effectiveness of current social welfare policies and hence make better judgments.

2 Introduction

Family, as the foundation of society, makes unique and irreplaceable contributions to the health of our economy. Today's family, nonetheless, is becoming increasingly diverse due to changing living standards, diversifying cultures and improving technologies. How modern families and affiliated individuals feel about life under rapid societal changes has become a topic of our wonder.

Our goal is to **investigate the influential factors affecting a person's life satisfaction and the extent of their impacts**. This is done through organizing and analyzing data retrieved from the 31st cycle, focusing on families, of the General Social Survey at Statistics Canada. Details of the dataset will be introduced in the section that follows. In the dataset, we assumed for gender, education, living area, marital status, self-rated health and family income to be the correlated variables that will impact a person's life satisfaction as a whole and the result is surprising.

The sections that follow will elaborate on the basis of data selection, details of the regression model, regression results, discussion of results, model limitations and further improvements. Government bodies can consider the result in the policy-making process to determine which policies to adopt. Interested groups and individuals can use the result to evaluate the effectiveness of current social welfare programs.

3 Data

We obtained this dataset from the 31st cycle of the General Social Survey ("GSS") conducted by the Diversity and Sociocultural Statistics at Statistics Canada. This dataset contains 461 variables, derived from the responses on a telephone-based Questionnaire collected from 20,602 respondents.

The **sampling method** of this dataset is unique and complex, with stratification, multiple stages of selection and unequal selection probabilities for respondents. To more detail, the **target population** is all persons fifteen years of age and older in Canada, excluding the residents of Yukon, Northwest Territories and Nunavut, as well as full-time residents of institutions. A total of 27 strata were formed by geographic areas, in which 17 are Census Metropolitan Areas ("CMAs") and the remaining are non-CMAs.

The **frame population**, on the other hand, is created with two different components. The first one being the list of telephone numbers in use, both landline and cellular, available to Statistics Canada. The latter would be the Address Register, a list of all dwellings within the ten provinces of interest, used to group together all telephone numbers associated with the same valid address. The **sampling frame** [1] is hence the combination of telephone numbers and the Address Register.

The dataset has several **advantages**. Firstly, it comprises a large pool of data, collecting beyond its desired sample size, with a variety of attributes concerning facets relevant to a family navigating through conjugal, family, and work trajectories. Moreover, the responses and information contained in this data set are also very much recent, collected and carefully weighted to represent the entire population. Specifically, each

[1] Upon collection of responses, the telephone numbers would be grouped into records that consist of the sampling unit on the survey frame. Then, each record in the frame was assigned to a stratum within its province. Lastly, a simple random sample without replacement of records was performed in each stratum.

person selected in the sample represents several other persons not in the sample, and the number of persons represented by a given person is determined using a weighting factor of the sampled person. Therefore, this dataset should be a good representation of both the sampled and non-sampled population.

However, some **weaknesses** are evident too. The most apparent ones being the sampling and non-sampling error. **Sampling error**, the difference between the estimates obtained from the sample and the results from a complete count taken under similar conditions, appears because the survey is designed to be based on a sample of persons. That is, have the same design and methodology be applied to a complete census, the results might be different. **Non-sampling errors**, on the other hand, can happen at every stage too [2]. However, the most prominent source of non-sampling errors would be non-responses on survey results, both partial and total. The former occurs when respondents fail to answer one or a few questions, whereas the latter occurs when the respondents fail to answer all questions. This may lead to the volunteer bias where the voluntary participants in the study do not represent the entire population. As discussed in the GSS31 user guide published by Statistics Canada, non-respondents tend to be younger males. The non-responses and the resulting unequal distribution of demographic characteristics are handled by adjusting the weight of households who responded to the survey to cover for those who did not.

Moreover, aside from the sampling and non-sampling errors, the choice of frame population may limit the target population's ability to represent. As previously stated, the frame population is created through the telephone numbers available to Statistics Canada. Neglecting the fact that there may be contact numbers available to Statistics Canada already, households without telephones were excluded from the survey population. One may argue that this population is minimal, but this conclusion cannot be drawn unless further studies on the demographic group without telephones are done.

Additionally, as discussed previously, this survey has its unique and complex design, which could affect the estimation and variance calculations used in common analyses, posing analytical challenges. Also, since the responses from this survey are based on a sample of persons, different figures might be obtained if a complete census had been taken using the same Questionnaire, interviewers, supervisors, processing methods, etc.

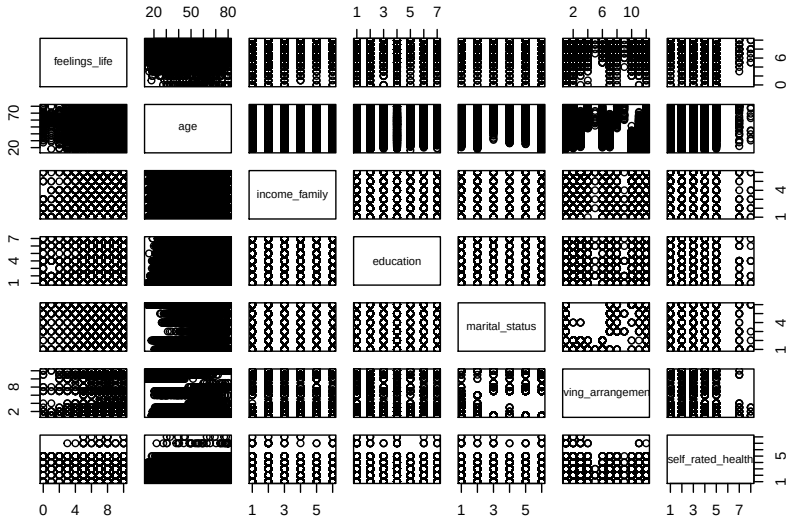


Figure 1: Plot of raw data

The raw data was initially filtered, reorganized, and renamed by gss_cleaning.r (Rohan & Sam Caetano, 2020), and the total variable dropped from 461 to 81 by cleaning. Highly correlated variables and variables with high proportions of missing values were excluded during the cleaning process to avoid multicollinearity

[2]Examples include interviewers misunderstanding instructions, respondents making mistakes in answers, errors appearing in inputting, processing and tabulating data, and etc.

and to increase model efficiency.As shown in the scatter plot (Figure 1), the majority of responses are coded by scale and they are discrete variables.

The questionnaire is detailed and some variables can overlap each other. We pay more attention to variables on family level rather than individual level. In the end, we narrowed our scope to six independent predictor variables of interest. They are gender, education, living area, marital status, self-rated health and family income. After omitting missing values within this scope, the final dataset used for regression analysis contains 19,873 observations.

4 Model

The most of variables in the dataset is discrete and lack linearity between covariates and response which violate assumption of simple linear regression. Though a linear regression model may be more straightforward and easier to analyse, it lacks linearity between covariates and response and violates the assumption of linear model. Therefore, we have chosen to use logistic regression instead, which is more suitable when analysing multi-characteristic discrete variables.

The life satisfaction responses collected in the GSS (2017) were individuals' self-rated feelings on a scale of 0 to 10, and the average is relatively high. It is unclear how one unit change of life satisfaction will impact a person's behaviours, and we have We have hence divided the life feeling responses by their mean value (8.09), to differentiate between individuals with above-average life satisfactions (indicated by 1) and those without (indicated by 0). The resulting new variable, ***Life_Satisfaction***, which is an indicator, is the response variable of our model.

Under our hypothesis, an improvement in education level, family income, marital status or self rated health is set to increase the odds of life satisfaction above average. Male is assumed to be more likely to have above average life satisfaction than female, due to gender inequalities at workplace, home, etc. Individuals living in rural areas or Prince Edward Island are more likely to have above average life satisfaction than those living in urban areas, due to a more relaxed lifestyle.

The basis of our variable selection is as follows: For our logistic model, "sex" variable and "is_male" variable do not lead to statistical difference. We selected "sex" for a more straightforward intepretation. We assume that an individual's appetite for a living environment (pop_center) impacts more on his or her life satisfaction than the general geographic region (province or region). We consider the special cases such as underaged individuals taking income from their family, who may have above average life satisfactions with no personal income. "Income_respondent" is hence removed from our model. We assume that an individual's physical health will have greater influence on his or her life satisfaction. "self_rated_mental_health" was also removed to avoid multicollinearity. "marital_status" is chosen over "living_arrangement" since it is more descriptive and specific. For instance, by living "along" under "living_arrangement", an individual can be separated, divorced or widowed.

Some common demographic characteristics such as age and minority status are not selected due to lack of representative. Others including average hours worked and occupation are excluded from the regression model due to high proportions of missing values.

Then, our basic expression of our model is shown as following:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

where p represents the probability to have life satisfaction above average. x_i(i=1,. . . ,n) correspond to levels of *gender*, *education level*, *living area*, *family income*, *marital status* and *self_rated_health*. (Thus, n = sum of level for each predictor has - number of predictors, since one level will used as based line i logistic model), β_i(i=2,..,7) are estimate coefficient, β_0 is intercept, ϵ is random error.

For better representation of data and comprehensive interpretation of the model, we have simplified the classification of certain categorical predictor variables by merging their affiliated categories into new categories. Categories are aggregated as referenced in Appendix A: Summary of Category Aggregations (Table 4).

Table 1: Summary of Model

	Estimated	Confidence	Interval	P value
(Intercept)	1.608	1.379	1.875	1.47e-09
sexMale	0.925	0.87	0.983	1.20e-02
educationLess than high school diploma or its equivalent	1.437	1.299	1.589	1.83e-12
educationBachelor or above	0.831	0.763	0.906	2.46e-05
educationcollege / below bachelor level	0.937	0.865	1.016	1.16e-01
pop_centerRural (non CMA/CA)/Prince Edward Island	1.178	1.096	1.265	8.33e-06
marital_statusLiving common-law	1.23	1.065	1.421	4.77e-03
marital_statusMarried	1.754	1.557	1.977	2.71e-20
marital_statusSeparated	0.63	0.505	0.785	3.85e-05
marital_statusSingle, never married	0.754	0.665	0.855	1.01e-05
marital_statusWidowed	1.31	1.132	1.516	2.84e-04
self_rated_healthFair	0.159	0.141	0.181	9.64e-175
self_rated_healthGood	0.359	0.333	0.386	2.93e-160
self_rated_healthPoor	0.102	0.082	0.126	1.23e-96
income_family$25,000 to $49,999	0.799	0.724	0.88	6.14e-06
income_familyLess than $25,000	0.721	0.639	0.812	8.71e-08
income_family$50,000 to $124,999	0.887	0.822	0.958	2.25e-03

However, we do recognize that the results and outputs from a logistic regression model is harder to interpret and analyze. The predictions would need to be inferred from odds and exponents, which will be done in the following sections.

All analysis of data is done by Rstudio software.

5 Result

The final model result indicates that gender, education, living area, marital status, self-rated health and family income are the significant (p value < 0.05) predictors of life satisfaction. In addition, self-rated health, some of education, marital status and family income levels are especially significant ($p<0.001$). The response of logistic model is log odds as described in last section, we did exponential transfer for the estimated coefficient to order to more direct interpretation. The estimated coefficient after exponential calculation, confidence interval, and p-value are summarized in Table 1.

In addition, the AUC (area under curve) = 0.68 as the ROC curve in appendix (Figure 2) illustrates that the predictive of the final model is 0.68, which is considered acceptable. Also, as shown in appendix Table 2, the model build use test dataset gets similar coefficients of the model as when using the training dataset, so the final model is validated through cross-validation. That is, the model is reasonable valid and predictive.

6 Discussion

As discussed in previous sections, the data is obtained from the 31st cycle of the General Social Survey ("GSS") conducted by the Diversity and Sociocultural Statistics at Statistics Canada. The volunteer bias may exist and reweighting was performed to adjust for that. Response variable is also transformed to categorical variable to avoid the unequal distribution of original variable caused by such bias.

According to the summary (Table 1), the odds of life satisfaction level above average will decrease by about 10% for males than females, increase by 17.8% for individuals living in rural areas or Prince Edward Island than those living in urban or CMA/CA areas. Self rated health is shown as highly significant in determining

the admit. Compare with *Excellent*, the odds of life satisfaction level above average will decrease by 60%, 84%, 90% correspond to is *Good*, *Fair*, *Poor*.

Compared with high school and its equivalent education level, the odds of more life satisfaction will increase by 43.7% when education level is less than high school, decrease by 16.9% when education level is bachelor or above. The confidence interval of the coefficient of college / below bachelor level education contains 1 which indicates that there is no significant difference.

As well, compared with family income more than $125,000, the odds of above average life satisfaction will decrease by 28%, 20%, 12% when family income is Less than $25,000, $25,000 to $49,999, $50,000 to $124,999, respectively.

In contrast with marital status' baseline value *divorce*, the odds of more life satisfaction will decrease by 37%, 24.6% when the respondent's marital status is *separated*and *single*, and increase by 23%, 31%, 75.4% when it is *living common law*, *widowed*, and *married*.

The results indicate that living area, family income and self-rated health comply with our hypothesis, whereas gender, education and marital status have brought us surprises.

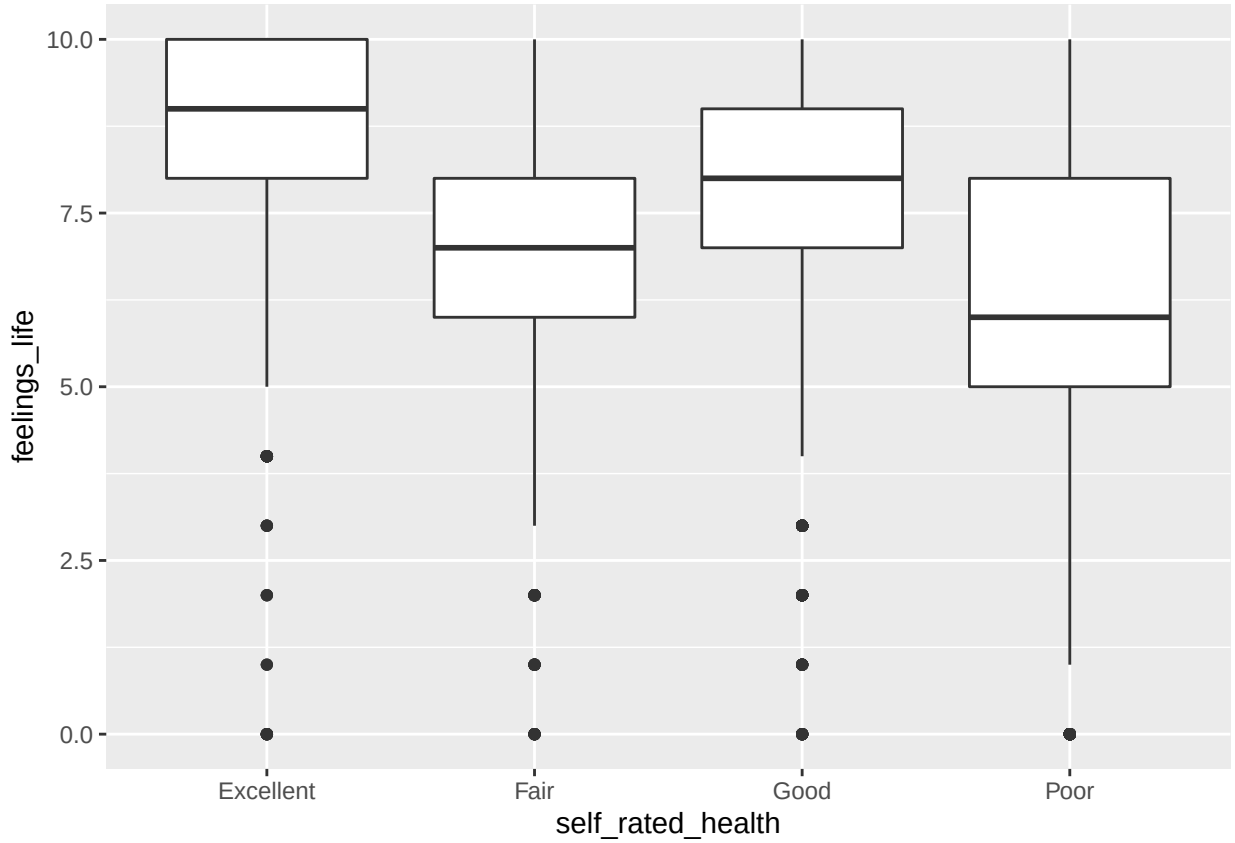
As implied by the results, the increase in family income and the improvement of self-rated health can make an individual more probable to have above average life satisfactions. People living in rural areas or Prince Edward Island are more likely to have outstanding life satisfactions, in contrast with those living in large urban areas.

However, the higher an individual's education level is, the less likely he or she will have life satisfactions above average. This is possibly due to the social structure of Canada. As a developed country, Canada has more skilled workers than unskilled workers. Such imbalance has caused an excess supply of skilled workers and an excess demand for unskilled workers, which has resulted in the competitive job market for undergraduates and high labor cost.

Furthermore, females are slightly more likely to have above average life satisfactions than males. The odds will increase by a small difference for females than males. This could possibly imply that the issue of gender inequality is contained well in Canada.

Lastly, marital status complies partially with our expectations. People married or living common-law are generally more likely to have above average life satisfactions than people divorced and single or separated people are generally less likely to have above average life satisfactions. However, it is out of our expectations that widowed people have higher odds than the divorced. We have found no good explanation for this phenomenon. It is hard to argue whether "widowed" is an improvement from "divorced".

The model can provide reference to government bodies on the effectiveness of current social welfare policies. For instance, self-rated health is an extremely significant variable which brings significant difference among each of its affiliated categories. This could imply that Canadian social welfare policies should focus on Canadians' health and the universal healthcare program is potentially a great contributor to Canada's high average life satisfaction. As well, government bodies, educational institutions and interested individuals can analyze the results on education coefficients to reconsider the necessity of education from a demographic perspective instead of a financial perspective. Universities can also reconsider the justifiability of increasing tuitions.



As indicated by the sample boxplot (Figure **??**) that measures the relationship between feelings_life and self_rated_health, there are apparent differences between each health category. A clear trend of increasing life satisfaction exists as the respondents' health condition improves.

6.1 Weaknesses

The analysis performed has a few apparent weaknesses.
Firstly, and probably the most apparent, the mean of both variable Feeling_life and Age is somewhat distorted towards the higher end. Note that the variable Feeling_life is ordinal in the dataset, ranging from 0 to 10 inclusive, and its mean is 8.09 when split as indicator variable, leaning toward the higher end. This either suggests that the sampled population has a very positive feeling towards their lives, or that this variable was over-reported during the survey process and cannot be used to represent and predict the population.

The high age mean, on the other hand, suggests that the majority of people who responded to the survey are of the older age groups. Therefore, the analysis performed on such respondent demographic composition could have limited representation power on the entire target population.

Secondly, as discussed in the previous sections, we omitted numerous variables that have high correlation or high proportions of missing values to avoid multicollinearity and to increase model efficiency. However, some variables that were omitted may be potentially effective and would be logical to consider when evaluating a person's life satisfaction level. For example, average number of working hours, mental health status and occupation. Excluding them may seem beneficial from the model building perspective, but illogical otherwise.

6.2 Next Steps

The following paragraphs will offer some insights into potential next steps for future analyses.
As previously discussed, the frame population of the survey that provided information for the dataset is created with two components, in which the major one is the list of telephone numbers available to Statistics

Canada. By the nature of such a frame population, households without a telephone number available to Statistics Canada are omitted from the survey population, hindering the representation ability of the sampled population. Therefore, we would suggest conducting a follow-up survey using different approaches to cover as much of the target population as possible. An alternative approach could be a computer-assisted personal interviewing (CAPI), or computer-assisted self-interviewing (CASI), both of which would allow a broader exposure to the target population via digital technology. The use of digital technology may also enable the follow-up survey to reach younger respondents, as the age mean of the survey presented in the dataset is somewhat distorted towards the higher end. Note that the follow-up surveys could narrow the general survey topic down to analysis-specific ones to increase the amount of relevant data available for the analysis.

Additionally, other data analysis techniques could be used. For example, software like Python could be used in addition to R, as done in previous sections, to allow for a more concise and detailed analysis. Also, data could be inputted using automated machines to reduce manual input errors.

7 References

1. Alexander, R., & Caetano, S. (2020, 10 07). gss_cleaning.R.
2. Beaupre, P. (2020). General Social Survey Cycle 31: Families, Public Use Microdata File Documentation and User's Guide. Ottawa: Authority of the Minister responsible for Statistics Canada.
3. Statistics Canada. 2017. Census Profile. 2016 Census. Statistics Canada Catalogue no. 98-316-X2016001. Ottawa. Released November 29, 2017. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E
4. T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0.
5. Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/
6. Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.
7. Ethan Heinzen, Jason Sinnwell, Elizabeth Atkinson, Tina Gunderson and Gregory Dougherty (2020). arsenal: An Arsenal of 'R' Functions for Large-Scale Statistical Summaries. https://github.com/mayoverse/arsenal, https://cran.r-project.org/package=arsenal, https://mayoverse.github.io/arsenal/.

8 Appendix

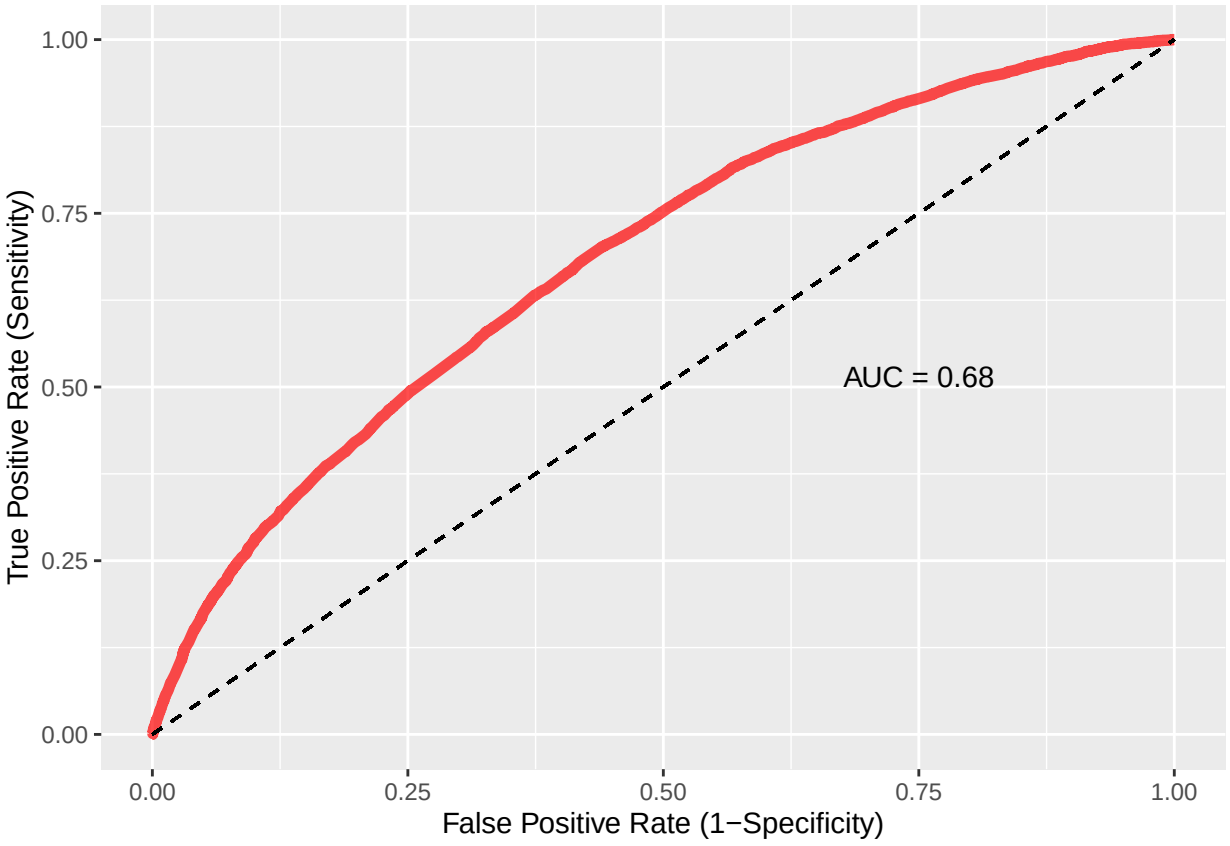


Figure 2: ROC curve

Table 3: Summary of Data

	Atlantic region (N=4564)	British Columbia (N=2522)	Ontario (N=5621)	Prairie region (N=4073)	Quebec (N=3822)	Total (N=20602)	p value
age							< 0.001
Mean (SD)	53.526 (17.244)	53.052 (17.868)	51.913 (17.822)	50.746 (18.059)	51.974 (17.681)	52.190 (17.747)	
Range	15.200 - 80.000	15.000 - 80.000	15.000 - 80.000	15.000 - 80.000	15.000 - 80.000	15.000 - 80.000	
sex							0.132
Female	2537 (55.6%)	1352 (53.6%)	3082 (54.8%)	2157 (53.0%)	2075 (54.3%)	11203 (54.4%)	
Male	2027 (44.4%)	1170 (46.4%)	2539 (45.2%)	1916 (47.0%)	1747 (45.7%)	9399 (45.6%)	
feelings_life							0.007
N-Miss	67	30	85	53	36	271	
Mean (SD)	8.161 (1.635)	8.023 (1.684)	8.064 (1.696)	8.101 (1.678)	8.101 (1.514)	8.094 (1.645)	

	Atlantic region (N=4564)	British Columbia (N=2522)	Ontario (N=5621)	Prairie region (N=4073)	Quebec (N=3822)	Total (N=20602)	p value
Range	0.000 - 10.000	0.000 - 10.000	0.000 - 10.000	0.000 - 10.000	0.000 - 10.000	0.000 - 10.000	
place_birth_canada							< 0.001
N-Miss	18	13	29	24	13	97	
Born in Canada	4237 (93.2%)	1757 (70.0%)	3806 (68.1%)	3300 (81.5%)	3255 (85.5%)	16355 (79.8%)	
Born outside Canada	277 (6.1%)	752 (30.0%)	1780 (31.8%)	738 (18.2%)	550 (14.4%)	4097 (20.0%)	
Don't know	32 (0.7%)	0 (0.0%)	6 (0.1%)	11 (0.3%)	4 (0.1%)	53 (0.3%)	
vis_minority							< 0.001
N-Miss	25	23	43	33	16	140	
Don't know	53 (1.2%)	32 (1.3%)	123 (2.2%)	67 (1.7%)	39 (1.0%)	314 (1.5%)	
Not a visible minority	4376 (96.4%)	1994 (79.8%)	4265 (76.5%)	3504 (86.7%)	3435 (90.3%)	17574 (85.9%)	
Visible minority	110 (2.4%)	473 (18.9%)	1190 (21.3%)	469 (11.6%)	332 (8.7%)	2574 (12.6%)	
citizenship_status							< 0.001
N-Miss	114	202	356	294	177	1143	
By birth	4252 (95.6%)	1779 (76.7%)	3826 (72.7%)	3318 (87.8%)	3264 (89.5%)	16439 (84.5%)	
By naturalization	166 (3.7%)	539 (23.2%)	1429 (27.1%)	453 (12.0%)	377 (10.3%)	2964 (15.2%)	
Don't know	32 (0.7%)	2 (0.1%)	10 (0.2%)	8 (0.2%)	4 (0.1%)	56 (0.3%)	
own_rent							< 0.001
N-Miss	27	18	33	29	13	120	
Don't know	27 (0.6%)	16 (0.6%)	15 (0.3%)	11 (0.3%)	2 (0.1%)	71 (0.3%)	
Owned by you or a member of this household, even if it i...	3467 (76.4%)	1817 (72.6%)	4168 (74.6%)	3162 (78.2%)	2480 (65.1%)	15094 (73.7%)	
Rented, even if no cash rent is paid	1043 (23.0%)	671 (26.8%)	1405 (25.1%)	871 (21.5%)	1327 (34.8%)	5317 (26.0%)	
income_family							< 0.001
$100,000 to $ 124,999	434 (9.5%)	267 (10.6%)	634 (11.3%)	464 (11.4%)	359 (9.4%)	2158 (10.5%)	
$125,000 and more	898 (19.7%)	608 (24.1%)	1510 (26.9%)	1036 (25.4%)	655 (17.1%)	4707 (22.8%)	
$25,000 to $49,999	1050 (23.0%)	515 (20.4%)	1024 (18.2%)	821 (20.2%)	935 (24.5%)	4345 (21.1%)	
$50,000 to $74,999	844 (18.5%)	419 (16.6%)	993 (17.7%)	715 (17.6%)	725 (19.0%)	3696 (17.9%)	
$75,000 to $99,999	679 (14.9%)	366 (14.5%)	778 (13.8%)	559 (13.7%)	539 (14.1%)	2921 (14.2%)	

	Atlantic region (N=4564)	British Columbia (N=2522)	Ontario (N=5621)	Prairie region (N=4073)	Quebec (N=3822)	Total (N=20602)	p value
Less than $25,000	659 (14.4%)	347 (13.8%)	682 (12.1%)	478 (11.7%)	609 (15.9%)	2775 (13.5%)	
education							< 0.001
N-Miss	71	42	90	80	58	341	
Bachelor's degree (e.g. B.A., B.Sc., LL.B.)	717 (16.0%)	520 (21.0%)	1179 (21.3%)	687 (17.2%)	650 (17.3%)	3753 (18.5%)	
College, CEGEP or other non-university certificate or di...	1068 (23.8%)	519 (20.9%)	1388 (25.1%)	836 (20.9%)	755 (20.1%)	4566 (22.5%)	
High school diploma or a high school equivalency certificate	1065 (23.7%)	620 (25.0%)	1302 (23.5%)	1082 (27.1%)	779 (20.7%)	4848 (23.9%)	
Less than high school diploma or its equivalent	804 (17.9%)	269 (10.8%)	660 (11.9%)	582 (14.6%)	721 (19.2%)	3036 (15.0%)	
Trade certificate or diploma	354 (7.9%)	173 (7.0%)	184 (3.3%)	356 (8.9%)	416 (11.1%)	1483 (7.3%)	
University certificate or diploma below the bachelor's level	135 (3.0%)	122 (4.9%)	158 (2.9%)	178 (4.5%)	139 (3.7%)	732 (3.6%)	
University certificate, diploma or degree above the bach...	350 (7.8%)	257 (10.4%)	660 (11.9%)	272 (6.8%)	304 (8.1%)	1843 (9.1%)	
pop_center							< 0.001
Larger urban population centres (CMA/CA)	2592 (56.8%)	2246 (89.1%)	5067 (90.1%)	2898 (71.2%)	3124 (81.7%)	15927 (77.3%)	
Prince Edward Island	708 (15.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	708 (3.4%)	
Rural areas and small population centres (non CMA/CA)	1264 (27.7%)	276 (10.9%)	554 (9.9%)	1175 (28.8%)	698 (18.3%)	3967 (19.3%)	
marital_status							< 0.001
N-Miss	1	0	1	4	1	7	
Divorced	350 (7.7%)	228 (9.0%)	478 (8.5%)	333 (8.2%)	378 (9.9%)	1767 (8.6%)	
Living common-law	443 (9.7%)	195 (7.7%)	381 (6.8%)	258 (6.3%)	798 (20.9%)	2075 (10.1%)	
Married	2185 (47.9%)	1243 (49.3%)	2708 (48.2%)	2125 (52.2%)	1240 (32.5%)	9501 (46.1%)	
Separated	183 (4.0%)	79 (3.1%)	206 (3.7%)	95 (2.3%)	80 (2.1%)	643 (3.1%)	
Single, never married	903 (19.8%)	562 (22.3%)	1328 (23.6%)	888 (21.8%)	1029 (26.9%)	4710 (22.9%)	
Widowed	499 (10.9%)	215 (8.5%)	519 (9.2%)	370 (9.1%)	296 (7.7%)	1899 (9.2%)	
living_arrangement							< 0.001
Alone	1247 (27.3%)	747 (29.6%)	1544 (27.5%)	1097 (26.9%)	1180 (30.9%)	5815 (28.2%)	

	Atlantic region (N=4564)	British Columbia (N=2522)	Ontario (N=5621)	Prairie region (N=4073)	Quebec (N=3822)	Total (N=20602)	p value
Living with one parent	109 (2.4%)	49 (1.9%)	160 (2.8%)	81 (2.0%)	112 (2.9%)	511 (2.5%)	
Living with two parents	167 (3.7%)	121 (4.8%)	337 (6.0%)	188 (4.6%)	159 (4.2%)	972 (4.7%)	
No spouse and non-single child(ren)	1 (0.0%)	2 (0.1%)	9 (0.2%)	5 (0.1%)	3 (0.1%)	20 (0.1%)	
No spouse and single child 25 years of age or older	41 (0.9%)	22 (0.9%)	78 (1.4%)	41 (1.0%)	31 (0.8%)	213 (1.0%)	
No spouse and single child under 25 years of age	190 (4.2%)	70 (2.8%)	238 (4.2%)	130 (3.2%)	186 (4.9%)	814 (4.0%)	
Other living arrangement	254 (5.6%)	153 (6.1%)	292 (5.2%)	269 (6.6%)	166 (4.3%)	1134 (5.5%)	
Spouse and non-single child(ren)	1 (0.0%)	1 (0.0%)	7 (0.1%)	2 (0.0%)	0 (0.0%)	11 (0.1%)	
Spouse and other	46 (1.0%)	21 (0.8%)	69 (1.2%)	44 (1.1%)	25 (0.7%)	205 (1.0%)	
Spouse and single child 25 years of age or older	91 (2.0%)	40 (1.6%)	137 (2.4%)	61 (1.5%)	44 (1.2%)	373 (1.8%)	
Spouse and single child under 25 years of age	799 (17.5%)	482 (19.1%)	1191 (21.2%)	890 (21.9%)	774 (20.3%)	4136 (20.1%)	
Spouse only	1618 (35.5%)	814 (32.3%)	1559 (27.7%)	1265 (31.1%)	1142 (29.9%)	6398 (31.1%)	
self_rated_health							< 0.001
N-Miss	19	16	33	19	12	99	
Don't know	18 (0.4%)	6 (0.2%)	18 (0.3%)	10 (0.2%)	5 (0.1%)	57 (0.3%)	
Excellent	833 (18.3%)	566 (22.6%)	1206 (21.6%)	839 (20.7%)	932 (24.5%)	4376 (21.3%)	
Fair	542 (11.9%)	248 (9.9%)	517 (9.3%)	426 (10.5%)	345 (9.1%)	2078 (10.1%)	
Good	1303 (28.7%)	771 (30.8%)	1676 (30.0%)	1250 (30.8%)	1162 (30.5%)	6162 (30.1%)	
Poor	216 (4.8%)	116 (4.6%)	234 (4.2%)	155 (3.8%)	95 (2.5%)	816 (4.0%)	
Very good	1633 (35.9%)	799 (31.9%)	1937 (34.7%)	1374 (33.9%)	1271 (33.4%)	7014 (34.2%)	
self_rated_mental_health							< 0.001
N-Miss	19	19	34	22	12	106	
Don't know	16 (0.4%)	4 (0.2%)	16 (0.3%)	9 (0.2%)	12 (0.3%)	57 (0.3%)	
Excellent	1156 (25.4%)	714 (28.5%)	1726 (30.9%)	1134 (28.0%)	1350 (35.4%)	6080 (29.7%)	
Fair	340 (7.5%)	174 (7.0%)	361 (6.5%)	261 (6.4%)	160 (4.2%)	1296 (6.3%)	
Good	1385 (30.5%)	730 (29.2%)	1504 (26.9%)	1196 (29.5%)	998 (26.2%)	5813 (28.4%)	
Poor	79 (1.7%)	54 (2.2%)	99 (1.8%)	70 (1.7%)	24 (0.6%)	326 (1.6%)	

	Atlantic region (N=4564)	British Columbia (N=2522)	Ontario (N=5621)	Prairie region (N=4073)	Quebec (N=3822)	Total (N=20602)	p value
Very good	1569 (34.5%)	827 (33.0%)	1881 (33.7%)	1381 (34.1%)	1266 (33.2%)	6924 (33.8%)	
religion_has_affiliation							< 0.001
N-Miss	48	52	84	71	27	282	
Don't know	39 (0.9%)	35 (1.4%)	32 (0.6%)	42 (1.0%)	11 (0.3%)	159 (0.8%)	
Has religious affiliation	3840 (85.0%)	1494 (60.5%)	4386 (79.2%)	3049 (76.2%)	3386 (89.2%)	16155 (79.5%)	
No religious affiliation	637 (14.1%)	941 (38.1%)	1119 (20.2%)	911 (22.8%)	398 (10.5%)	4006 (19.7%)	
children_in_household							< 0.001
No child	3678 (80.6%)	2015 (79.9%)	4323 (76.9%)	3039 (74.6%)	2942 (77.0%)	15997 (77.6%)	
One child	433 (9.5%)	236 (9.4%)	577 (10.3%)	421 (10.3%)	378 (9.9%)	2045 (9.9%)	
Three or more children	92 (2.0%)	70 (2.8%)	184 (3.3%)	212 (5.2%)	131 (3.4%)	689 (3.3%)	
Two children	361 (7.9%)	201 (8.0%)	537 (9.6%)	401 (9.8%)	371 (9.7%)	1871 (9.1%)	
number_marriages							< 0.001
Mean (SD)	0.845 (0.607)	0.860 (0.651)	0.820 (0.611)	0.842 (0.606)	0.626 (0.610)	0.799 (0.620)	
Range	0.000 - 4.000	0.000 - 4.000	0.000 - 4.000	0.000 - 4.000	0.000 - 4.000	0.000 - 4.000	

Table 2: Cross Validation of Model

	test	train	P value
(Intercept)	1.447	1.724	8.031e-08
sexMale	0.909	0.936	1.023e-01
educationLess than high school diploma or its equivalent	1.382	1.483	3.322e-09
educationBachelor or above	0.808	0.843	2.687e-03
educationcollege / below bachelor level	0.958	0.925	1.441e-01
pop_centerRural (non CMA/CA)/Prince Edward Island	1.172	1.178	4.891e-04
marital_statusLiving common-law	1.372	1.15	1.401e-01
marital_statusMarried	1.972	1.63	4.126e-10
marital_statusSeparated	0.684	0.594	3.245e-04
marital_statusSingle, never married	0.849	0.698	1.294e-05
marital_statusWidowed	1.331	1.295	6.432e-03
self_rated_healthFair	0.163	0.156	3.862e-108
self_rated_healthGood	0.367	0.353	5.487e-100
self_rated_healthPoor	0.098	0.104	8.468e-60
income_family$25,000 to $49,999	0.825	0.779	1.159e-04
income_familyLess than $25,000	0.716	0.722	3.990e-05
income_family$50,000 to $124,999	0.841	0.918	9.127e-02

Table 4: Summary of Category Aggregations

Variable Names	**Original Categories**	**Aggregated Categories**
Gender (“sex”)	Male	Male
	Female	Female
Education Level (“education”)	Bachelor’s degree	Bachelor or above
	University certificate, diploma or degree above the bach...	
	University certificate, diploma or degree below the bachelor’s level	college/below bachelor level
	College, CEGEP or other non-university certificate or di...	
	Trade certificate or diploma	
	High school diploma or a high school equivalency certificate	High school diploma or a high school equivalency certificate
Living area (“pop_center”)	Rural areas and small population centres (non CMA/CA)	Rural areas (non CMA/CA) or Prince Edward Island
	Prince Edward Island	
	Larger urban population centres (CMA/CA)	Larger urban population centres (CMA/CA)
Family income (“income_family”)	Less than $25,000	Less than $25,000
	$25,000 to 49,999	25,000 to $49,999
	$50,000 to 74,999	$50,000 to 124,999
	$75,000 to 99,999	
	$100,000 to 124,999	
	$125,000 and more	above $125,000
Marital Status (“marital_status”)	Single, never married	Single, never married
	Married	Married
	Living common-law	Living common-law

	Separated	Separated
	Widowed	Widowed
	Divorced	Divorced
Self Rated Health (“self_rated_health”)	Excellent	Excellent
	Very good	Good
	Good	
	Fair	Fair
	Don’t know	
	Poor	Poor