DEBUNKING THE SIX FIGURE INCOME PROMISE

Cesar Y. Villarreal Guzman

October 19, 2020

Abstract

Based on the recent social media advertisement trend of financial courses, we investigate the likelihood of having an annual income higher than or equal to $100,000CAD given young adult males who are self employed and work on average less than 30-40 hours a week. To accomplish this we fitted an elastic net logistic regression model using the Canadian 2016 General Social Survey and concluded on the unlikeliness of this event.

Introduction

The year 2015 marked the beginning of the boom of the online education industry, but more specifically financial education. The famous advertisement video "Here In My Garage" by Tai Lopez,[1] which at the moment has nearly 71 million views, displayed the immense potential of social media advertisement as a viable marketing strategy. This particular video created a movement of the so called "Financial Freedom" courses, which allegedly offer a formula, a recipe, to not only making a six figure income, but to do so by being self employed and working less time per week than the average.

We would like to point out that this kind of "promises" are in fact not a new trend, there are many books on this topic which where published long before 2015. Some known authors include, Robert Kiyosaki and Tim Ferriss. In these past years we have seen new competitors enter this "Financial Freedom" market via YouTube and other social media platforms, we gave the example of Tai Lopez, but some other examples are Alex Becker and Dan Lok.

In this investigation we answered the question: How likely are you to make a six figure income in Canada? Moreover we discuss on the validity of these "Financial Freedom" courses, since our results show that it is more likely that a young adult gets an acceptance offer to Harvard University than to make a six figure income. Moreover we are also interested in the number of worked hours per week, since most of these courses claim that it is possible to have a high income while working 4-20 hours per week. To be clear the range 4-20 is an average of many different advertisements on this topic.

Although there is no available evidence, one can make an educated guess based on YouTube's viewership demographics that most of these ads are being targeted to young adult males,

[1] Link to video: https://www.youtube.com/watch?v=Cv1RJTHf5fk&ab_channel=TaiLopez

and hence the reason why we focus on this gender for our discussion. However the model is fitted to work on any age group or gender.

To generate results that help answer our question, we look at the General Canadian Social Survey (Statistics Canada 2017) the 2016 version, which had as focus "Canadians at Work and Home". We elaborate more on this data set in the "Data" section. The main tool used for analysis in this investigation is the elastic-net logistic regression model.

Data

The General Social Survey (GSS), as per the "Users Guide", is a sample survey with cross-sectional design. The target population includes all persons 15 years of age and older in Canada, excluding residents of the Yukon, Northwest Territories, and Nunavut and full-time residents of institutions.[2] The survey uses a frame that combines telephone numbers with Statistics Canada's Address Register.

Sampling was done by dividing each of the ten provinces into strata with many of the census metropolitan areas considered as separate strata. Each record in the survey frame was assigned to a stratum within its province. Then from each stratum a selection of records was selected by simple random sampling. During collection, households that did not meet the eligibility criteria were terminated after an initial set of questions. If the household was eligible then a respondent was randomly selected from each household to complete an electronic questionnaire or to respond to a telephone interview.

The GSS aimed at collecting 20,000 samples but only achieved a sample size of $n = 19,609$. The overall response rate was 50.8%, a very high response rate for a survey of this nature. Responses where adjusted using a weighting strategy which was followed by standard bootstrap weighting.

The study, was divided into a first set of questions that had the purpose of demographic classification, questions like age, sex, marital status, relationship of household members to respondent etc... The second section asked questions about life at work, life at home, work life balance and health, well-being and resilience. Income data was obtained through a linkage to tax data for respondents who did not object to it.

Data was retrieved from the Computing in the Humanities and Social Sciences (CHASS) at the University of Toronto website. A subset of the data was selected and retrieved for this investigation. This subset included demographics, as well as income and work related variables. The resulting data set contained 94 columns/features and 19609 observations. The data was cleaned by selecting wanted variables and remapping code values to text. Variables used for this report were: sex, age group, average number of work hours per week, income, and self employment. More information on how to retrieve this data set and the code used to clean it can be found in the Appendix. Figures 1 and 2 display the raw data used for the creation of the model.

[2]Households without telephones were excluded from the survey population.

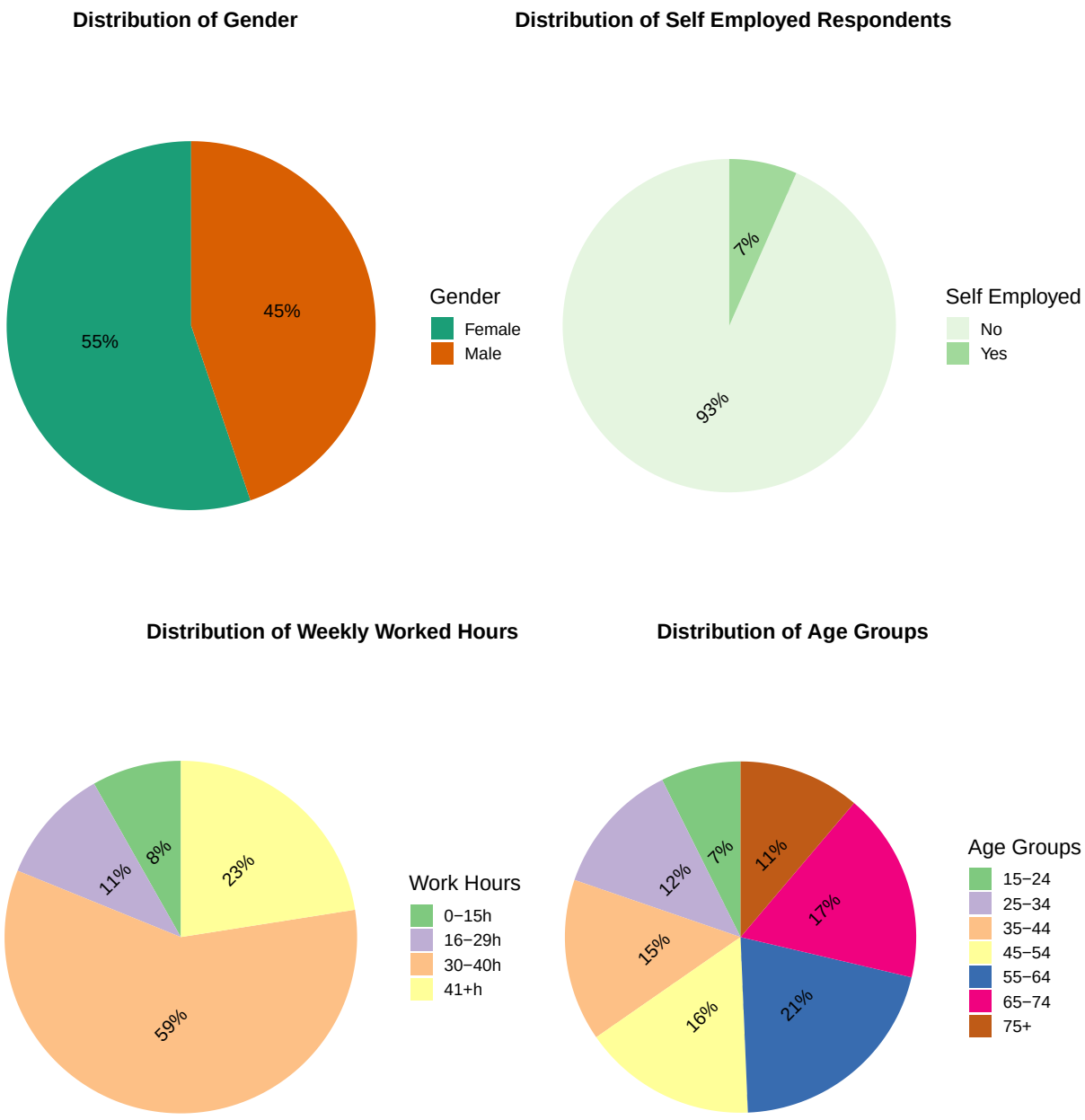

Figure 1: GSS Distributions of Predictor Variables

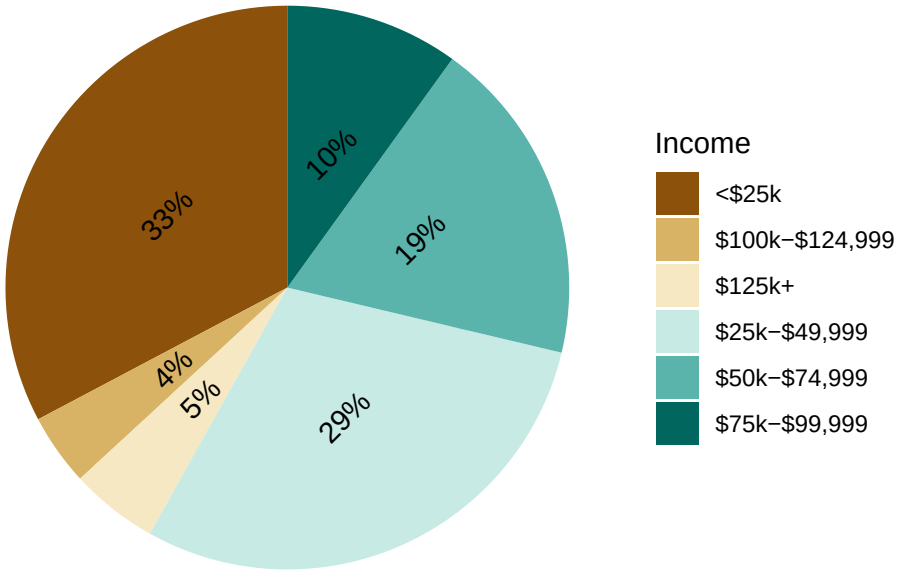

Figure 2: GSS Distributions of Income.

Model

In the introduction we outlined that we seek to model the likelihood of someone having a six figure income based on sex, age, number of hours worked per week and, if this individual is self employed. Fortunately, the GSS data set contains all of these variables except for age. To solve this issue we use the age group of respondents. Moreover, a new vector $\mathbf{y} = (y_1, \ldots, y_n)$ was generated for which $y_i = 1$ represents that respondent i has an income higher than or equal to \$100,000CAD and $y_i = 0$ otherwise. From this point on we will denote $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ as the matrix were observation $\mathbf{x}_j = (x_{j1}, \ldots, x_{jp})^T$ has p predictor features. In this particular case we have $p = 4$ predictor features.

The original question is then simplified to the problem of predicting the binary value y given $\mathbf{x}$. An implicit assumption of a linear relationship is made here to go along the research question and the opening discussion of the paper. Therefore, our model of choice, which is perhaps the most used and suited for binary classification, is the logistic regression model.

Logistic regression arises from the desire to model the posterior probability of the two classes 0 and 1 via linear functions. Suppose that the probability of having a six figure income is equivalent to p, then the model has the form

$$\log \frac{p}{1-p} = \beta_0 + \beta^T \mathbf{x}$$

here we assume $\beta = (\beta_1, \ldots, \beta_4)^T$ is a vector of appropriate size for left multiplication with

x. Logistic regression models are usually fit by maximum likelihood.(Hastie, Friedman, and Tisbshirani 2017)

In particular we use a regularized version of the logistic regression model. This is because, generally speaking regularization tends to be beneficial for the predictive performance of a model. There are three common methods for regression regularization: lasso (Tibshirani 1996), ridge (Hoerl and Kennard 1970) and elastic net regression (Zou and Hastie 2005).

Given strong prior knowledge it may be better to pick lasso or ridge in place of elastic net. However, in absence of prior knowledge, elastic net should be the preferred solution and our regularization method for this problem. The limitation here is that elastic net has significantly higher computational complexity and is therefore much slower than lasso or ridge. Both ridge and lasso represent viable alternatives to elastic net regularization and for the model as a whole.

Elastic net regression combines both strengths and solves the limitations of lasso and ridge regression. This is achieved by combining the L_1 and L_2 penalties. Following the notation used above, the log-likelihood for the conventional logistic regression model for N examples is given by

$$\ell(\beta_0, \beta) = \sum_{i=1}^{N} y_i \log p + (1 - y_i) \log(1 - p)$$

$$= \sum_{i=1}^{N} y_i(\beta_0 + \beta^T \mathbf{x}_i) - \log(1 + exp\{\beta_0 + \beta^T \mathbf{x}_i\})$$

then, the elastic net logistic regression model has the following log-likelihood for N examples

$$\ell_{elnet}(\beta_0, \beta) = \left[\sum_{i=1}^{N} y_i(\beta_0 + \beta^T \mathbf{x}_i) - \log(1 + exp\{\beta_0 + \beta^T \mathbf{x}_i\})\right] - \left[\lambda_1 \sum_{j=1}^{4} |\beta_j| + \lambda \sum_{j=1}^{4} \beta_j^2\right]$$

in the equation above we assume that $\beta = (\beta_1, \ldots, \beta_4)$. Also in this new equation, the value $\lambda_1 \sum_{j=1}^{4} |\beta_j|$ is the L_1 regularization penalty, and similarly $\lambda \sum_{j=1}^{4} \beta_j^2$ the L_2 regularization penalty. We use this new function to compute the pair $(\hat{\beta}_0, \hat{\beta})$ given by

$$(\hat{\beta}_0, \hat{\beta}) = \arg\max_{(\beta_0, \beta)} \ell_{elnet}(\beta_0, \beta)$$

Or equivalently

$$(\hat{\beta}_0, \hat{\beta}) = \arg\min_{(\beta_0, \beta)} -\ell_{elnet}(\beta_0, \beta)$$

Typically algorithms that solve this optimization problem tend to converge since log-likelihood is concave.

We use the implementation of this model found in the R package **glmnet** (Friedman, Hastie, and Tibshirani 2010). For the training phase we used a 80-20 train to test ratio split for the GSS data set. That is 80% of the data was used to fit the model and the remaining 20% was used to evaluate the performance of the model. Since all variables are categorical, we required the creation of several dummy binary variables, also known as "one-hot" encoding.

Results

Table 1: Coefficient Values for Fitted GSS Model

Variable Name	Coefficient
(Intercept)	-7.00
resp_age_group25-34	2.75
resp_age_group35-44	3.64
resp_age_group45-54	3.98
resp_age_group55-64	4.05
resp_age_group65-74	3.64
resp_age_group75+	4.04
sexMale	0.74
nhours_pweek16-29h	0.06
nhours_pweek30-40h	0.90
nhours_pweek41+h	1.67
nhours_pweekVS	-0.44
self_emplYes	-0.24

Table 1 displays the fitted model for our classification problem. It is important to note that Table 1 does not include any standard errors of regression coefficients or other estimated quantities. This is because as Friedman, Hastie, and Tibshirani (2010) explain in their implementation notes "... this package deliberately does not provide them. The reason for this is that standard errors are not very meaningful for strongly biased estimates such as arise from penalized estimation methods."

The accuracy on the test set was of **70.56%**, with sensitivity **74.86%** and specificity **70.13%**. Sensitivity, in this problem, measures the proportion of high income individuals that are correctly identified, and specificity measures the proportion of non-high income individuals that are correctly identified. We also include a confusion matrix plot, Figure 3, of our model on the test set which can be found in the "Discussion" section.

Discussion

In statistical terms, Figure 3 serves as evidence that our model lacked precision because of a high number of false positives. However, our achieved model accuracy is much better than chance which is closer to 50%.

With regards to the coefficients, it is no surprise that the most significant values are for the older age groups. It is reasonable to assume that as one gets older the chances of having a higher position at a company, or simply having profitable investments is higher. It is also no surprise that there is a positive relationship between number of hours and income. In a country like Canada where jobs are usually compensated based on hourly work this is what's to be expected.

What's interesting, however, is the little effect of the self employment variable on this model. The fact that the self employed group are a minority in most societies displays itself in this

Figure 3: Confusion Matrix of GSS Model on Test Set.

result.

Some of the weaknesses of our approach come from the limitations of using linear models on an apparent non-linear problem. As an alternative, we suggest fitting a support vector machine model. This method has the benefit of being partially linear but breaking its linearity with what is referred to as the "kernel trick". Another alternative is to fit a decision tree classifier.

Similarly, a limitation of our approach is that we do not posses any data on entrepreneurs who fit the demographic profiles. Some areas of future work are to sample a population of entrepreneurs and ask questions regarding the amount of time it took to start generating a six figure income, assuming they already make six figures. We focused on young adult males leaving behind other genders. Another area of future work could be to apply the same research question to females and compare the results.

Finally, as per the introduction, consider a young adult male in the age group 25-34 who is self employed and who works less than the average of 30-40 hours. Our model then estimates that the probability of having a six figure income is of **2.18%**, a very unlikely outcome. On the contrary we can say that our model shows evidence that the probability of having a high income is increased given an older person, working more than the average and not being self employed.

Appendix

Data was retrieved from the Computing in the Humanities and Social Sciences (CHASS) at the University of Toronto website. In order to retrieve this data set from the CHASS website, a valid University of Toronto email is required. Other academic institutions offer this data set but instructions will not be included.

The official link to the CHASS website is **http://dc.chass.utoronto.ca/myaccess.html**, on arrival to the site click on the option, located on the side bar, "SDA @ CHASS". Login with your University of Toronto credentials. Select your language of choice and on the keyboard click "Ctrl + f", Command for Mac users, and type "gss". Hit enter and click on "General Social Survey". Locate the 2016 version and click on "data", then on the top bar, under "Download" click "Customized Subset". We recommend downloading a "csv" file together with the codebook. Our cleaning code does not require any data definitions file. To ensure our code works we advise to select "All" variables. Variables of interest can be selected instead of "All" however our code may not work. At the bottom click continue and download all files.

All cleaning and model code used for this project can be found here:

https://github.com/cesar-yoab/elnet_regression_R

References

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. http://www.jstatsoft.org/v33/i01/.

Hastie, Trevor, Jerome Friedman, and Robert Tisbshirani. 2017. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hoerl, Arthur E., and Robert W. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12 (1): 55–67. https://doi.org/10.1080/00401706.1970.10488634.

Kuhn, Max. 2020. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Statistics Canada. 2017. "General Social Survey, Cycle 30: 2016: Canadians at Work and Home." http://www.chass.utoronto.ca/.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40 (1): 1–29. http://www.jstatsoft.org/v40/i01/.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wilke, Claus O. 2019. *Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'*. https://CRAN.R-project.org/package=cowplot.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.