National Graduates Survey 2018 Income Classification

AITAZAZ GILANI, University of Saskatchewan, Canada

ACM Reference Format:
Aitazaz Gilani. 2023. National Graduates Survey 2018 Income Classification. 1, 1 (October 2023), 2 pages.

1 INTRODUCTION

The National Graduates Survey (NGS) is a survey that collected information from persons who graduated from post-secondary institutions in Canada in 2015. The survey was taken in 2018, two years after students had graduated from their programs. Graduates were asked questions focused on their academic path, funding, and transition into the labour market. The data in the survey presents an opportunity to study the relationship between academic performance and involvement with employment opportunities and income. This project hopes to propose and implement a model that will help understand which significant factors during a post-secondary program can lead a student to a successful high-income career.

2 IMPORTANCE

The goal of this model is to help post-secondary institutions improve their programs to lead their graduates into successful, high-income careers. The model can also provide insight into market trends, such as which programs are valued higher in the labour market or whether certain skills or factors can lead a graduate to a high-income career. Students and academic advisors will also find value in the model in understanding a student's projected career income and what to focus on during their program.

3 CHALLENGES

The survey dataset only represents a small portion of graduates; only 19,565 graduates were interviewed out of 424,453 [STATSCAN 2018]. This leads to a sampling error present in the dataset, and generalizing the outcomes of the model to all graduates can have its issues. Non-sampling error is also present in the dataset; graduates could have made mistakes in interpreting or answering survey questions. Some graduates have also declined to answer survey questions important for this project. A significant portion of graduates have also abstained from answering their personal incomes post-graduation. Surveys missing important data points will need to be removed from the dataset before training and testing the model, further decreasing the sample size.

4 MODEL DOMAIN

The domain of the model falls under social economics, as it hopes to understand what factors can lead a student in a post-secondary institution to a high-income career. The model can be generalized to work with any survey data that contains labour market data, as the model's main goal is to understand what factors are significant in leading to a high-income career.

4.1 Income classification

The model will classify income levels based on categorical and quantitative variables such as grades, skills, program study, and other variables from their program. It will make use of a multinomial logistic regression model using MLE, as the outcome variable is a range of income levels. Logistic is used for classification as it shows the probability of an independent variable leading to a range of income-level classes.

4.2 Dependent and Independent variables

The independent variables include all survey variables under the Graduates 2015 program. These variables can range from categorical to quantitative. Examples of independent variables include work placement in 2015, entrepreneurial skills, grade in the program, program study, etc. Variables that contain personal or financial information such as gender, race, program funding, and debt will be omitted as this study assumes these are not significant factors in leading to a high-income career in a perfect world.

The dependent variable will be 10 categorical income levels ranging from \$ 10,000 to \$ 90,000. Survey participants who abstained from answering income questions in the survey will not be included in the training of the model to avoid problems.

5 INTERPRETING MODEL RESULTS

The model will use a 20–80 test train split. The accuracy of the model will be judged with the testing data, which can help gauge its correctness. Log likelihood will also be used to understand the goodness of fit; models with higher log likelihood and accuracy will be preferred. Independent variables that have a <0.05 P value will be seen as statistically significant contributors to increasing the probability of having a high-income career.

The independent variables must be meaningful and should not have a high correlation with another independent variable. If such a feature is present, it will be removed, and the model will be tested again for accuracy and the log likelihood to gauge for better outputs. The goal of the model is to find independent variables under <0.05 P value score and their beta coefficients. The coefficients will tell us what impact they have on the overall model, do they increase or decrease probability of reaching a high-income career?

REFERENCES

STATSCAN. 2018. National Graduates Survey - Public Use Microdata File, 2018. (2018). https://doi.org/10.25318/81M0011X-eng

Author's address: Aitazaz Gilani, University of Saskatchewan, Canada.

XXXX-XXXX/2023/10-ART \$15.00
https://doi.org/

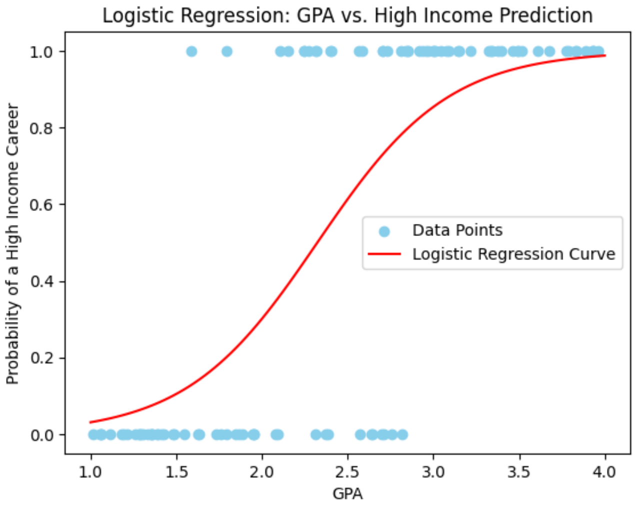

Fig. 3. Example simple figure showing probability of reaching a high income career with graduate GPA. The figure shows a logistic regression curve of one independent variable only and income classified as High and Low. Since the model will use a multinomial logistic regression with multiple categorical income level classifications, the following is a simplified visualization of the model.

A FIGURES

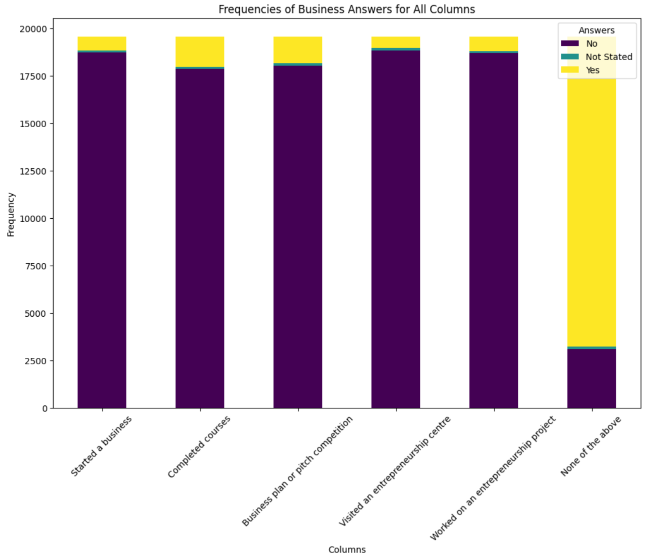

Fig. 1. Example data figure showing frequencies of answers on entrepreneurial skills demonstrated during graduates 2015 program. Bar graph shows a subset of the survey dataset showcasing responses from 6 questions from many graduates.

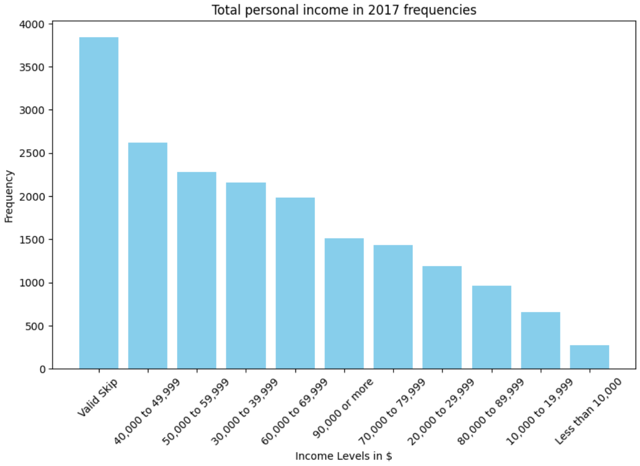

Fig. 2. Example dataset figure showing frequencies of answers on persons total personal income in 2017 post graduation. Bar graph shows responses for one question taken from all applicants.