# National Graduates Survey 2018 Income Classification Methodology

AITAZAZ GILANI, University of Saskatchewan, Canada

**ACM Reference Format:**
Aitazaz Gilani. 2023. National Graduates Survey 2018 Income Classification Methodology. 1, 1 (December 2023), 3 pages.

## 1 DATASET DESCRIPTION

The dataset provided for National Graduates Survey 2018 includes a cleaned CSV containing 124 columns depicting survey questions and 19,564 recorded encoded responses to the survey questions. This study is concerned with analyzing the relationship between the graduates 2015 program and their income in 2017.

### 1.1 Feature Selection

Variables that are relevant during their 2015 program will be chosen for this study as well the income variable in 2017. Remaining variables such as questions for prior to their 2015 program or post program will not be factored into the study. The dependent variable for this study is 'Total Personal income in 2017', to avoid inaccuracies all respondents who skipped the question will not factored into the dataset. This accounts for 13.3% of all survey respondents. The model will utilize a 20/80 percent test train split to evaluate for accuracy. Independent variables whose respondents also skipped may also be removed for the dataset depending on the leftover sample size of the income classes. Additional techniques such as PCA or dimensionality reduction might be used to help mitigate such issues.

### 1.2 Data pre-processing

The independent variables prior to training will be scaled using Z-Score normalization to help address multicollinearity. Variance inflation factor (VIF) will be used to rank independent variables, only selected variables under 5 (moderate multicorr.) will be factored into training. The dependent variable will also be re-adjusted to reduce the number of classes presented to 3 from 10 income classes denoted as 'Below average', 'Average income' and 'Above average' for a new graduate. This is done to present a more relevant and meaningful outcome and to address under sampling for certain classes. 'Average' under the income classes is defined as being in range of $50,000 - $59,999.

## 2 MULTINOMIAL LOGISTIC MODEL

The study will use a multinomial logistic regression to classify 3 income levels based of the normalized independent variables. Statsmodels and Scikit-learn will be the main libraries used to create the model. The parameters of the model by default will use 'multinomial' with additional parameters such as algorithm used or penalty will be evaluated based on accuracy of the model. Additional weight's might be applied to account for under sampling of the 'Average' income class depending on the f1-score of class. Several combinations of the model's parameters will be evaluated to find the most accurate and well performing model for all classes.

### 2.1 Interpreting model results

The goal of the study is to understand which variables are statistically significant in leading to an below average, average and above average income for a new graduate. Variables with P-values < 0.05 will be seen as statically significant for a particular income level. The coefficients of the model will also be important in understanding on what magnitude do they increase probability of reaching an income level.

### 2.2 Evaluating model

The model will mainly be evaluated on accuracy and log-likelyhood. Accuracy will be used to gauge for overfitting of the data and correctness. The model will be tested with many different combinations of paramters to test for the highest accuracy and f1-score per each outcome class. Log-likelyhood will also be used to understand goodness of fit, with higher log-likelyhood preferred with different combinations of parameters.

## 3 CHALLENGES

Normalization of the independent variables creates difficulty in interpreting their beta coefficients in the results. Since all the variables in the survey data are categorical, normalizing them losses important information regarding them. Normalization is needed to address multicollinearity however other strategies such as PCA might be used to create indices for a set like variables to mitigate this. However with normalized data, statistically significant variables per income class can still be found and interpreted. Small sample size of the 'Average' income class also remains a challenge as the model will struggle to create an a good fit for it. Class weights will be applied to penalize the model for misclassification of the 'Average' income class.

### 3.1 Sources of error

The results of the model and study will be affected by the sampling and non-sampling error present in the survey. The survey only interviewed a small sample size of gradutes and this number is further reduced in the training dataset as respondents who skipped answering their income in 2017 are removed. Non-sampling error is also present in the dataset as participants could have misinterpreted a question or have answered it wrong.

## A FIGURES

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


XXXX-XXXX/2023/12-ART $15.00
https://doi.org/

Table 1. Revised dataset table (21 features chosen out of 124)

| Feature | Encoding | Frequency | Training Frequency | Testing Frequency |
|---|---|---|---|---|
| Full-time or part-time student during 2015 program | PGM 034 | 17036 | 13628 | 3408 |
| Had a work placement during 2015 program | PGM 100 | 17036 | 13628 | 3408 |
| Worked during 2015 program | PGM 290 | 17036 | 13628 | 3408 |
| Volunteer activities during 2015 program | PGM 350 | 17036 | 13628 | 3408 |
| Program included components taken outside of Canada | PGM 380 | 17036 | 13628 | 3408 |
| Program taken through distance education | PGM P400 | 17036 | 13628 | 3408 |
| Overall grade average | PGM P405 | 17036 | 13628 | 3408 |
| Program taken towards certificate, diploma or degree since 2015 | EDU 010 | 17036 | 13628 | 3408 |
| Number of programs taken since 2015 program | EDU 020 | 17036 | 13628 | 3408 |
| Employee or self-employed last week | LFW P140 | 17036 | 13628 | 3408 |
| Job last week permanent or not permanent | LFW 270 | 17036 | 13628 | 3408 |
| Region of educational institution for 2015 program | REG INST | 17036 | 13628 | 3408 |
| Level of study for 2015 program | CERTLEVP | 17036 | 13628 | 3408 |
| Aggregated CIP 2016 for 2015 program | PGMCIPAP | 17036 | 13628 | 3408 |
| Respondents who participated in a co-op program | COOP | 17036 | 13628 | 3408 |
| Highest level of education completed at time of graduation | HLOSGRDP | 17036 | 13628 | 3408 |
| Total amount received from scholarships/awards/fellowships and prizes | SCHOLAR | 17036 | 13628 | 3408 |
| Debt size of all loans at time of graduation | DBTALGRD | 17036 | 13628 | 3408 |
| Total personal income in 2017 | PERSINCP | 17036 | 13628 | 3408 |
| 2015 level of education relative to father's level of education | FATEDGRD | 17036 | 13628 | 3408 |
| 2015 level of education relative to mother's level of education | MOTEDGRD | 17036 | 13628 | 3408 |

A.1

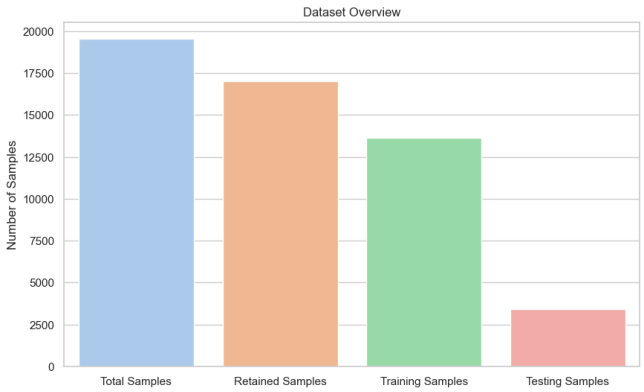


Fig. 1. Bar graph showing the number of samples in the original, filtered, training and testing datasets. Original dataset contained 19564 samples which were filtered to 17036 responses. This was done by removing participants who skipped answering 'Total personal income in 2017'. Figure also show cases the test-train split from the retained dataset kept after filtering, an 20/80 percent test-train split was used.

A.2

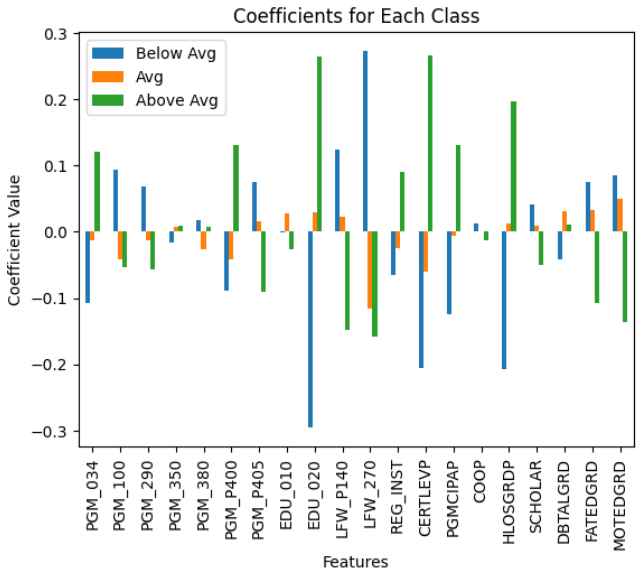


Fig. 2. Figure showcasing the coefficients results for a preliminary analysis conducted on a multinomial model using the mentioned dataset. Figure showcases the coefficients for the 3 income classes "Below Average", "Average" and "Above Average". Default parameters were used on the model and the mentioned data pre-processing steps were taken.

A.3

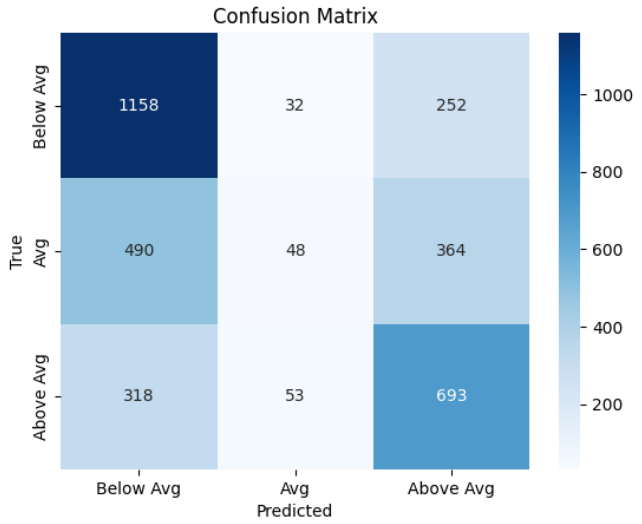


Fig. 3. Figure showcasing the confusion matrix of the preliminary analysis. The figure shows weakness in the model on predicting 'Average' income class due to a small sample size. The model showed an accuracy of 55% applying the above techniques and test train split with a weak f1-score of 0.09 for 'Average' income classification in regards to the other classes.