National Graduates Survey 2018 Income Classification Results

AITAZAZ GILANI, University of Saskatchewan, Canada

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

1 CHANGES IN METHODS

The updated model for the study utilizes a simple binary logistic regression from the previous multinomial model. This was changed to address an unbalanced dataset as there were not enough data points for 'Average' and 'Above average' income classes, resulting in poor accuracy even with weights applied. In the current model, 'Above average' is defined as total income $60,000 and above.

1.1 Feature Selection

Variables related to labour market status at the time of the interview were dropped from the dataset. The variables 'LFW P140' and 'LFW 270' were not included in the final dataset as they were not relevant to the participants educational backgrounds.

1.2 Data pre-processing

The dependent variable (total income in 2017) was updated to contain additional 1565 variables extracted from the 'JOBINCP' variable (salary of the last or current job). The variables were only extracted for participants who skipped answering the dependent variable 'PERSINCP' but had not skipped 'JOBINCP'. Additionally, the dataset was not normalized to address multicorrelinearity, as interpreting model coefficients becomes challenging.

2 MODEL RESULTS

The resulting model had an accuracy of 73% which was better than the preliminary multinomial model, which had an accuracy of 56%. Additionally, statistically significant predictors of reaching 'Above average' income were selected based on their P-values above 0.05 (A.3). There are 12 predictors out of 18 dependent variables that were chosen, with the most important ones having a positive coefficient. CO-OP programs and debt at graduation were not statistically significant variables as they are not important in determining an 'Above average' income level. The model also saw an improved log-likelihood of -8060.3 compared to -12768 in the multinomial model.

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

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

2.1 Interpreting coefficients results

To see how certain features such as grades (PGM P405) impact the likelihood of reaching an 'Above average' income, the odds ratio must be considered. The odds ratio for a selected answer can be defined as

$$e^{response \times coefficient}$$

An example would be if a participant picked response 1 for PGM P405 corresponding to A- to A+ grade, then given the coefficient, the odds ratio would be

$$e^{1 \times -0.1547}$$

Since the odds ratio for the following is under 1, we can say that it decreases the odds of reaching an 'Above average' income. This shows the significance of positive and negative coefficients for independent variables; the magnitudes and responses tell us the impact they have on having higher odds of reaching 'Above average' income.

2.2 Parameter and cross-validation testing

The model was tested for the solver used in logistic regression. The model was evaluated for its accuracy based on the solver after passing a 5-fold cross-validation test to get the mean accuracy for a given solver. On average, the solvers did not significantly overperform the default parameters, as seen in figure A.2. The solvers varied between 73% and 72%. We also see that changing the randomness of assigned variables in a 20/80 test train split for a cross-validation test did not significantly impact the model's accuracy, as seen in figure A.1. The model, on average, remains consistent at 73% accuracy.

3 CHALLENGES

A major challenge with the dataset is the imbalance; there are only 5746 values that are considered 'Above average' income for a graduate compared to the remaining 12859 values. The f1 score for classifying 'Above average' income was 0.42, compared to 0.82 for under 'Above average' income. The differences in scoring show that classifying 'Above average' income is harder for the model; applying weights to the model could improve performance, but it would come at the cost of decreasing f1 scoring for the other class. Another issue with the dataset is 'Skipped' responses are factored in for several independent variables, potentially skewing the results. They could not be factored out as removing skipped responses to decrease the dataset frequency by a large degree. Another issue is that participants who skipped answering total income but answered their job's salary may not be accurate as they may not factor in their other sources of income. The model assumes the job's salary and income for these participants must be at a similar level.

The results of the model are also affected by the sampling and non-sampling errors present in the survey. The survey only interviewed a small sample size of graduates (19,564), and this number is reduced by 959 participants who skipped answering their total income.

A FIGURES

Table 1. Revised dataset table (19 features chosen out of 124)

Feature	Encoding	Frequency	Training Frequency	Testing Frequency
Full-time or part-time student during 2015 program	PGM 034	18605	14884	3721
Had a work placement during 2015 program	PGM 100	18605	14884	3721
Worked during 2015 program	PGM 290	18605	14884	3721
Volunteer activities during 2015 program	PGM 350	18605	14884	3721
Program included components taken outside of Canada	PGM 380	18605	14884	3721
Program taken through distance education	PGM P400	18605	14884	3721
Overall grade average	PGM P405	18605	14884	3721
Program taken towards certificate, diploma or degree since 2015	EDU 010	18605	14884	3721
Number of programs taken since 2015 program	EDU 020	18605	14884	3721
Region of educational institution for 2015 program	REG INST	18605	14884	3721
Level of study for 2015 program	CERTLEVP	18605	14884	3721
Aggregated CIP 2016 for 2015 program	PGMCIPAP	18605	14884	3721
Respondents who participated in a co-op program	COOP	18605	14884	3721
Highest level of education completed at time of graduation	HLOSGRDP	18605	14884	3721
Total amount received from scholarships/awards/fellowships and prizes	SCHOLAR	18605	14884	3721
Debt size of all loans at time of graduation	DBTALGRD	18605	14884	3721
Total personal income in 2017	PERSINCP	18605	14884	3721
2015 level of education relative to father's level of education	FATEDGRD	18605	14884	3721
2015 level of education relative to mother's level of education	MOTEDGRD	18605	14884	3721

A.1

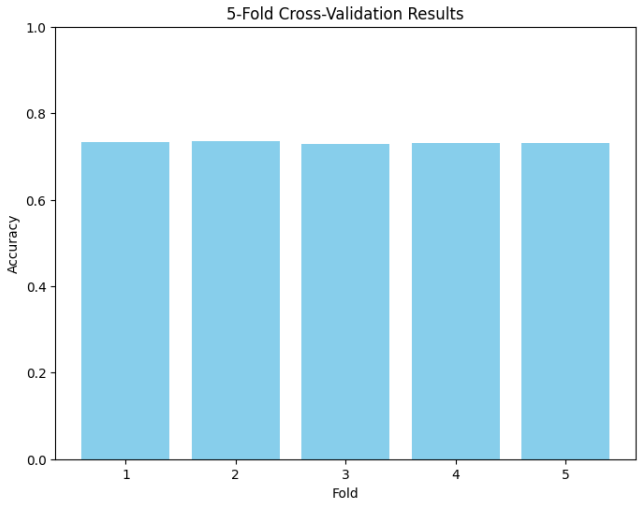

Fig. 1. Bar graph showing the results of doing 5-fold cross validation on default logistic regression parameters and the LBFGS solver. Each fold was evaluated for accuracy based on a randomized 80/20 test train split. The graph shows that, on average, the accuracy remains at 73% on each test.

A.2

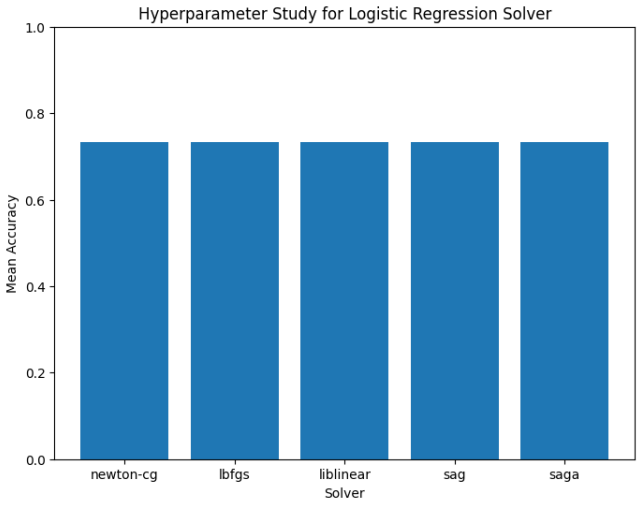

Fig. 2. Bar graph showing the results of parameter testing for the solver used in the model. Each bar represents the mean accuracy for a solver taken after conducting a 5-fold cross-validation test. On average, the solver did not make a difference in the model's accuracy, as it remains at 73%.

A.3

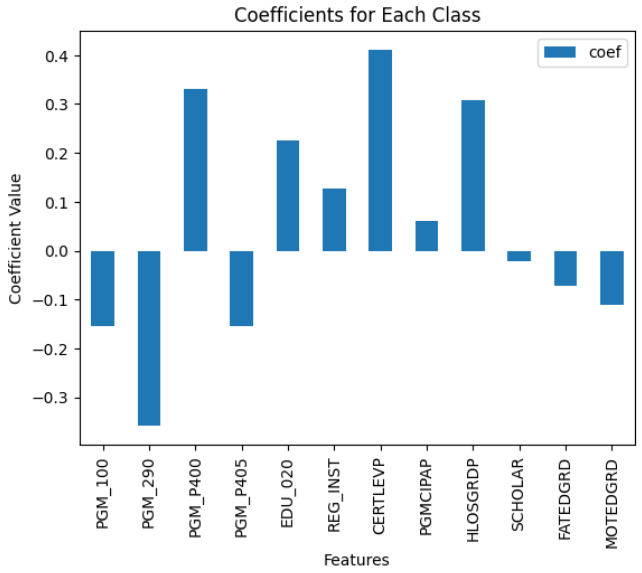

Fig. 3. Figure showcasing the coefficients for each statistically significant feature class in the model. The coefficients with the highest magnitude were Level of Study in 2015 (CERTLEVP), telling us the odds ratio for a given response will have a significant positive impact on reaching 'Above average' income. The lowest magnitude 'Worked during 2015 program' (PGM 290) tells us that for a given response, the odds ratio will be under 1, meaning a reduced likelihood of reaching 'Above average' income level.