Early Career Earnings of Post Secondary Graduates in Canada

Khashayar Zarekarizi, #1011627373

April 2024

JEL Codes: J24, J31, I21

Keywords: school-to-work transition, youth labour market outcomes, Earnings disparity

Abstract

This paper examined the impact of Education-related factors of graduates from Canadian Institutions(Diploma, Bachelors and Masters and Doctorate holders) to forecast their Income three years post-graduation by leveraging Canadian National Graduates Survey(NGS) 2013 and 2018 editions. Using Regression Analysis, found that variables such as GPA, Years of education, the reputation of the school attended and participation in a Co-op program have a statistically positive impact on average earnings. Additionally results differ by program studied and the field the individual is employed in, as engineering students and those in management occupations can expect higher incomes. Gender, Student-loans, and age also considered.

Introduction

It is widely a researched phenomenon that individuals with higher levels of education can expect to earn more from income(card 1999). In labour economics, the life-cycle agent model is widely used to explain the individual's choice of the optimal level of education based on their future earnings expectations(card 1999). The life cycle agent model is a framework used to analyze how individuals make decisions over their lifetimes regarding consumption, savings, and labor supply. It considers factors such as income, age, expectations about the future, and preferences. In the context of education choices, the life cycle agent model considers how individuals decide on investments in education over their lifetimes. This includes decisions such as whether to pursue higher education, what field of study to pursue, and how much to invest in education versus other expenditures. The model considers factors such as expected future earnings, the cost of education, and individuals' preferences for education and leisure.

In order to make the optimal decision about their educational needs, the individual needs to be given reliable information and estimates on how her decisions regarding educational factors such as field of study, Grade-point-average earned, participation in co-operative education program and the reputation of the institution attended along other factors which will impact their future incomes, specifically their post-graduate incomes in this case from three years after they have graduated their programs.

The purpose of this paper is to aim to make a generalizable model using the Canadian National Graduates Survey from the graduates in the year 2010 and 2015. Three regression

models were employed, the first was an OLS model on only the 2015 data that found variables such as GPA, Work placement and certification level, the program completed(engineering and architecture students had the highest expected incomes) the field they are occupied in as those occupied in managerial roles had the highest expected incomes.

The second model was a heteroskedasticity robust OLS model that used the combined dataset from 2010 survey with the 2015 survey and added a variable to account for differences between the year each person graduated.

When I combined the dataset from graduates of 2010 and 2015 I found that the graduates in 2010 that had been in the labour market from 2010 to 2013 had a higher average income ($58 127) compared to those who graduated in 2015 with an average income of ($56 347) three years after graduation. This result is even more emphasized since they are not adjusted for inflation yet as the income earned in 2013 would enable significantly higher consumption power than one earned in 2018. This shows the power of graduating during a time of economic easing versus the time of slow to moderate growth as those who graduated during 2010s were benefited with expansionary economic policy to aim the recovery from the 2008 crisis characterized by government stimulus packages and accommodative monetary policies which played a significant role in the early income of new graduates.

Literature Review

The Literature on the topic of economics of education is highly diverse and informative however there are not many academic papers trying to predict the outcomes of the income of graduates of post-secondary schools based on a high number of factors available. Most of the economics literature on the topic is trying to measure a causal impact of a particular variable such as grade point average, higher education, the student's socioeconomic background, gender, and other main variables of interest on the incomes of post secondary graduates and the covariables that they have used are mainly to assist with controlling for the heterogeneity between each person in the sample. However this paper's aim is to try to examine many variables that may impact the post graduate's early incomes both in conjecture with each other(through a large regression model) as well as trying to single out the impact of each significant variable.

To select variables for the model, hypothesis about their possible importance and finally to correctly interpret the results we had to rely heavily on papers that examined the independent variables I am using to the income output. Since these papers have a causal component to them, they are of great resource to interpret my results in this paper that aims to syndicate the results of the knowledge from these papers to build a robust model. Additionally, I used papers that used the same dataset as mine(the national graduates survey) to see how they used the dataset in their studies and finally I studied papers that used machine learning models to try to examine the question at hand to gauge the techniques used.

Finnie (1999) examined the outcome of the program field of study of bachelor's degree graduates across three cohorts of the national graduates survey. He had access to longitudinal

data from 1984 to 1995 editions of the database and follow up interviews. he studied the tangible results such as unemployment rates, income levels as well as job-education skill match and self reported job satisfaction. He found that majors that have a strong practical and skill-based part to them such as engineering and education tend to have high job placements, income, and overall satisfaction and those that graduated with more general degrees tend to have lower outcomes. His results are retested in this paper to see if the outcomes still holds after more than 20 years as I will be using program studied under the CIP 2016 classifications to understand the outcomes of graduates with respect to income levels.

However it is possible that some degrees attract individuals attract stronger students as they are harder to get into. Additionally higher abilities enables a person to accumulate more human capital as they enter a profession. This process will enable them to increase their incomes significantly faster than of those with lower abilities. (Boissiere 1985)

Another Paper that used the same dataset as this paper(NGS 2018) studies the impact of Grades on labour market income of graduates. (Mueller 2020) This suggests that higher grades are significant to achieving higher labour market earnings. This is in part attributed to the “signalling” effect that suggests employers perceive graduates with higher GPAs as more capable especially in fields such as math and business. However, work experience and additional education or training tend to somewhat mitigate these effects, suggesting that the strength of the grade average signal to employers is weakened. This study also found as the individual accumulates more human capital(through employment, upskilling, or farther education) the signalling effect of GPA becomes less important. This is understandable in the world of imperfect information where employers must make decisions without access to all the

information about potential candidates. This study used a Probit model while I am using an OLS model with different covariables that may result in different outcomes.

These results agree with (Roska 2010) which also found whilst the returns to higher education have been increasing with years, the variation in returns of different fields have widen greatly based on market demand of the skill-set of graduates.

After finding out that students have a lower income level in the 2015 data than the 2010 sample, the paper(Oreopoilos 2008) helped understand the rather unsuspecting result. In this paper the authors found that government and overall economic climate at the time of graduation has a highly important effect on the income of the new graduates. This which the authors called the “luck effect” predicts that students who graduate in times of easing economic policy characterized by low interest rates and unemployment tend to have higher income than the ones that graduate during recession.

(Francesconi 2018) examined the impact of early gender gaps among university graduates in Germany. They found that even though there are more women in university studies(like Canada) and they enter university with higher grades than men, they face a significantly lower income levels. This was aimed to be explained by field of study and grades as well as working hours and other measures of human capital accumulation and while the wage gap did decrease quite a bit after controlling for these factor there still remains an unexplained portion.

(Dearden 2019) was used to select the variable; institutional reputation in the regression model. This paper found that both to potential employers and to students, the reputation of the school attended is highly important, so important in fact that universities

have a large incentive to change their behaviour to maximize their school ranking among other schools in reputable rankings rather than maximizing student experience. The interpenetration of the institutional reputation on the income of graduates is thereby interesting since it will indirectly examine if ranking maximizing behaviour is positively correlated with the incomes of graduates from those schools. (Fitzgerald 2000) also examines the impact of college quality on earnings of recent grads found that attending a selective college or university were associated with an increase in earnings of 11-16 percent among men and women. However they also found that decisions made whilst in school such as field of study and grades have a combined higher impact on earnings.

(Rothstein 2011) examined the impact of student loans on graduates behaviours. Whilst they found that students with higher student debt levels tend to choose higher paying jobs, their income levels are adversely effected by their risk avoidance behaviour, lack of social connections(network) and unwillingness to partake in further education which goes to show that students with higher student loan balances tend to have lower early incomes.

(Wyonch 2020) found that the Coop program offered by universities in Canada have significant benefits for participants in the form of eased transition to the labor market and higher incomes after graduation and that they may play a role in overcoming wage gaps associated with bias toward individual characteristics (race,gender,immigration status). They also increase the probability that the student's first job after graduation will be in field closely related to his studies. The results of this study are re-examined especially since most coop students in the previous study were enrolled in a STEM program, not a uniform distribution of

students. This paper while using the same dataset, will use another more recent cohort for analysis.

(Biewen 2006) studied the impact of further training on employment in Germany, this paper uses methods such as perpensity score matching and other methods that may be usable in farther studies.

As a result of this literature review, I am hypothesizing that GPA, participation in a COOP program, reputation of the institution attended, the level of study(bachelors, college, masters or higher) have a positive impact on education and student loans levels and being a female is hypothesized to lower wages. Lastly the economic conditions in the period of job search, field of study and the industry occupied in are associated with deviations in income.

Data Description

The data used is from the national graduates survey of Canadian Graduates from 2013 and 2018 edition. This dataset is comprised of a comprehensive survey of graduates of Canadian educational institutes in college level, bachelors, and post bachelorette level. This questionnaire is distributed 3 years after the individual have finished their initial educational designation, this is significant as it means that the individual has had some time to settle into the labour market and find a job suitable to their perceived talents by the market. The Canadian National Graduates Survey (NGS) is a voluntary survey typically collected through a combination of survey methods such as online surveys, telephone interviews, and paper questionnaires. The survey gathers data from graduates across Canada, Once the data is collected, it undergoes a series of transformations including data cleaning, anonymization to ensure protecting personal identity of respondents, and weighting to reduce sampling bias.

This database collects information on the individual's demographic information, educational background and characteristics, financial information and assistance and employment status. To select a sample that is representative of the aim of the paper, I dropped any observation that is still enrolled in post secondary studies as well as anyone who is currently not employed. This is because to effectively study the impact of educational characteristics on current incomes the individual cannot be still a student nor unemployed. The final sample is representative of graduates of Canadian institutions without any age restrictions that are not currently enrolled in formal education and are employed(whether part-time or full-time).

The survey excludes graduates from private postsecondary institutions, graduates who completed continuing-education programs at colleges and universities that did not led to a diploma or degree, graduates in apprenticeship programs and graduates living outside of Canada at the time of the survey. I also deleted any miscoded observations or observations that had a null value for any of the variables. Since the dataset is very large we had the luxury of being able to work

To join the data from 2018 survey to the 2013 survey, there were wrangling and cleaning tasks required as the data and the questionnaire between 2013 and 2018 were different in many aspects. Tasks included recoding variables such as the occupation and program where the data within the variables did not match so I had to recode the data in stata to have a uniform encoding of the information. The way the person's grades were recorded was different between the two datasets; for the 2018 data the grades were chosen based on the letter grade received(A, B, C and D) but for the 2013 data, there was no matching variable however there was an Academic ranking variable that asked; in what percentage of your class

did you graduate(top 10%, 10-25%, lower than 25%, …). In the combined model both GPA and Academic ranking are used as independent variables and the differences between them are accounted using a (year graduated) control variable.

For the purpose of the study some imputed variables were created namely the institution reputation(inst_rep) which is a binary variable corresponding on whether thee student attended a highly reputable institution. This variable is imputed based on a question on the survey that asked; what was the most important factor in your choice of postsecondary institution? The students who answered that they chose the institution based on reputation were selected in the reputable (1) category and the rest in other(0) category. This factor was important to be considered since to protect the privacy of the respondents, the survey does not provide information on the school individuals attended, however the literature suggests that the place a person studied at has an important effect on their future earnings(Dearden 2019) so this variable was created as a proxy for the reputation effect of institution attended.

Another imputed variable is the year graduated binary variable which accounts for the year each observation graduated in the combined model to account for differences in graduating in different years.

Methodology

This study uses various methodologies to find the most appropriate results for the question at hand. The Main methodology used for this analysis are OLS regressions which is a useful method to find relationships between variables and interpret them confidently. To build up on that a Lasso regression model was used to normalize the data and help select the most impactful variables effecting post graduation wages. The results of this model are appropriate

for resource constraint individuals that do not have access to all the variables studied(for example they have not attended a reputable university) but still wish to maximize outcomes based on other factors in their disposal. Lastly a nonparametric K-nearest neighbors model was trained using the data to account for any variation not covered by the original OLS model.

The main dependent variable used for examining the outcome of student's post-graduation outcomes is their income level three years post graduation. This is a self reported income level earned by the individuals and is measured by the amount of Canadian dollars earned per year by the individual. To protect the privacy of the individuals the data in this PUMF are recoded based on a range they fall into(for example 1 represents individuals with 0-10000 of income and 9 represents people with 80000 to 90000 of yearly income).

The main independent variable used in this study are the university program studied, the field the individual is currently occupied in, the grade point average received in their studies, the individual's class rankings, whether they participated in a coop program, whether they attended a reputable institution, the level of studies graduated from(College, Bachelors, post-Bachelorette studies). The control variables used were the age of the respondent, gender, and the year they graduated to account for differences between graduation cohorts.

Individual's GPA is a significant factor in the literature on incomes of recent grads as they are used as a signal of the individual's abilities and competency to potential employers. This leaves individuals with an incentive of maximizing GPA by perhaps taking an easier course-load, majoring in lower difficulty fields and other ways instead of using the time in school to maximize human capital. Thereby the individual's GPA on their income makes for a good case study of early results of such behaviour. In this study GPA is measured as self reported grades.

In the 2013 Dataset there were no measures of GPA, so a Class ranking variable is used that asked students in what class ranking did they graduate(top 10%, 10-25%, lower than 25%) coded as 1-3 in the data. This is used as a proxy for grades effect in the 2013 data.

There was a variable used for participation in Co-op program which is a work placement program for students still enrolled in schools. This variable is a binary variable(0 for participants and 1 for non participants) indicating if they attended a coop program or not, as the literature supports that coop graduates can expect to earn higher incomes and an easier work-to-school transition than those who didn't partake.

The variable institution reputation is a derived variable based on if the individual attended a selective school or not, derived from those that said they attended their academic institution mainly because of the reputation of the school.

The certification level of the individual is also considered as the theory suggests that a rational actor would only partake in higher years of schooling(with its direct and indirect associated costs) in order to maximize labour market outcomes(which is of course not only associated with an individuals short term earnings only). This is examined through the certification level of individuals by asking what level of studies(bachelors, diploma, masters or doctorate) did they graduate from.

Additionally the program they studied was considered as some majors tend to have higher paying median incomes than others. This field differences was captured from their "Classification of Instructional Programs (CIP) Canada 2016 classification that was a little different than the classification used for the 2013 dataset but the differences were dropped when compiling the data between two datasets.

The field that the individual is currently employed in is also considered using the NOC 2016 classifications of industries. The data. Were again cleaned to be compatible to the 2013 data.

The first control variable used were the age of the person to account for the human capital accumulated outside of the individual's study. Since the dataset had no measures of individual's natural ability, a person's age is our best estimate of the accumulated human capital in this case. Ranges from(less than 25, 25-29, 30-39, 40 or more)

The second control variable is whether they took any student loans, it is a binary variable that asked of they took any government student loans? This is to account for individual's socioeconomic backgrounds as less financially advantaged students tend to apply for more student loans(Rothstein 2010) Lastly the Gender of the students were controlled for as there's evidence suggesting an early gender wage gap among graduates that sometimes persists, it's a binary variable between men and women.

Results

Summary Statistics: Below we will discuss the main findings of this study and how to interpret them. The first visualization presented is the summary statistics of the variables used in the study to best understand the results of the regressions by quickly looking at these results.

	index	count	mean	std	min	25%	50%	75%	max
0	income	19255.000000	5.727603	2.456078	1.000000	4.000000	5.000000	7.000000	10.000000
1	gpa_new	19255.000000	1.174396	1.295219	0.000000	0.000000	0.000000	2.000000	3.000000
2	class ranking	19255.000000	1.025811	1.141313	0.000000	0.000000	1.000000	2.000000	3.000000
3	age	19255.000000	1.930460	1.042413	1.000000	1.000000	2.000000	3.000000	4.000000
4	coop	19255.000000	1.875513	0.330145	1.000000	2.000000	2.000000	2.000000	2.000000
5	inst_rep	19255.000000	0.383641	0.486285	0.000000	0.000000	0.000000	1.000000	1.000000
6	student_loan	19255.000000	0.523345	0.499468	0.000000	0.000000	1.000000	1.000000	1.000000
7	gender	19255.000000	1.584212	0.492870	1.000000	1.000000	2.000000	2.000000	2.000000
8	year_grad	19255.000000	0.478110	0.499534	0.000000	0.000000	0.000000	1.000000	1.000000
9	occup_Business, finance and administration occupations	19255.000000	0.168060	0.373929	0.000000	0.000000	0.000000	0.000000	1.000000
10	occup_Health occupations	19255.000000	0.132693	0.339251	0.000000	0.000000	0.000000	0.000000	1.000000
11	occup_Management occupations	19255.000000	0.068865	0.253231	0.000000	0.000000	0.000000	0.000000	1.000000
12	occup_Natural and applied sciences and related occupations	19255.000000	0.158348	0.365077	0.000000	0.000000	0.000000	0.000000	1.000000
13	occup_Occupations in art, culture, recreation and sport	19255.000000	0.046585	0.210754	0.000000	0.000000	0.000000	0.000000	1.000000
14	occup_Occupations,education,law,social,community and gov services	19255.000000	0.250948	0.433570	0.000000	0.000000	0.000000	1.000000	1.000000
15	occup_Sales and service occupations	19255.000000	0.112283	0.315722	0.000000	0.000000	0.000000	0.000000	1.000000
16	occup_Trades,transport,equipment operators and related occupations	19255.000000	0.038276	0.191866	0.000000	0.000000	0.000000	0.000000	1.000000
17	program_Architecture, engineering, and related technologies	19255.000000	0.137315	0.344189	0.000000	0.000000	0.000000	0.000000	1.000000
18	program_Business, management and public administration	19255.000000	0.196313	0.397218	0.000000	0.000000	0.000000	0.000000	1.000000
19	program_Education	19255.000000	0.090158	0.286416	0.000000	0.000000	0.000000	0.000000	1.000000
20	program_Humanities	19255.000000	0.057336	0.232489	0.000000	0.000000	0.000000	0.000000	1.000000
21	program_Mathematics, computer and information sciences	19255.000000	0.046378	0.210307	0.000000	0.000000	0.000000	0.000000	1.000000
22	program_Other	19255.000000	0.066320	0.248848	0.000000	0.000000	0.000000	0.000000	1.000000
23	program_Physical and life sciences and technologies	19255.000000	0.056505	0.230900	0.000000	0.000000	0.000000	0.000000	1.000000
24	program_Social and behavioural sciences and law	19255.000000	0.144015	0.351114	0.000000	0.000000	0.000000	0.000000	1.000000
25	program_Visual and performing arts, and communications technologies	19255.000000	0.051104	0.220215	0.000000	0.000000	0.000000	0.000000	1.000000
26	cert_lvl_Bachelor's	19255.000000	0.434069	0.495647	0.000000	0.000000	0.000000	1.000000	1.000000
27	cert_lvl_College	19255.000000	0.303350	0.459717	0.000000	0.000000	0.000000	1.000000	1.000000
28	cert_lvl_Master's / Doctorate	19255.000000	0.256816	0.436889	0.000000	0.000000	0.000000	1.000000	1.000000
29	certification_level	19255.000000	1.941937	0.761501	0.000000	1.000000	2.000000	3.000000	3.000000

Notes- table 1: This table depicts each variable used in the analysis, it farther shows the mean value, the number of observations used(count), the degree of variation between results(std), the values each variable can take and the average percentile for each variable. We can see that

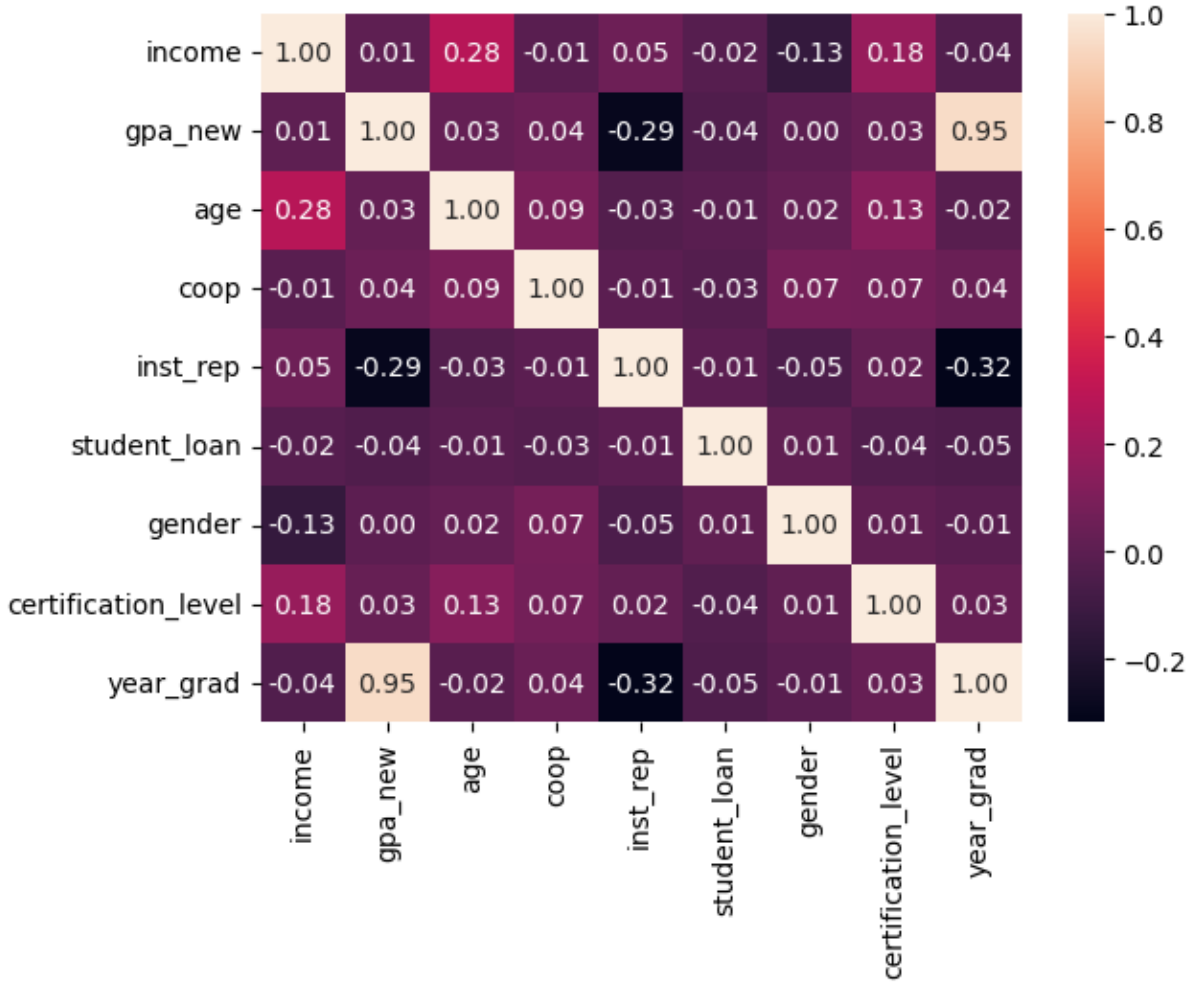
The mean income is 5.7276(corresponding to $57 276) and that can vary greatly based on the standard deviation of the observation. This paper aims to in part explain some of the large deviation in incomes based on the variables used. The average GPA is 1.174 which means its slightly lower than an A average. Class ranking reports a similar results but it's a bit rated higher. The age of graduation(1.93) suggests that the average graduate was 25-29 years old which is potentially skewed higher by masters and doctorate students.

Around 33% of the people participated in a coop program and 38% of the sample attended a school perceived as high reputation to them. The gender, student loans and year grad variables show that the observations are almost 50-50 distributed between those who took student loans and those who didn't, men and women and those who graduated in 2010 and ones graduated in 2015. Additionally the occupation and program categories were one-hot-encoded into separate binary values that show were most people at this age are occupied in a "education, law, social, community and gov services(25.1%)" category and "Business, finance and administration occupations(16.8%)". The majority of the people studied either Business(19.6%) or social sciences(14.4%) as their programs and the least popular program was "Mathematics, computer and information sciences" based programs with 4.6% of the observations. Lastly we see that the data is divided between 30.3% college graduates, 43.4% between bachelors degree holders and 25.7% by those holding a masters or doctorate degree.

In the next part, the relationship of these variables of interest are considered against each other as well as with our dependant variable(income) to find the best determinants of income of graduates of Canadian institutions.

Relationship between variables- Correlation

Heatmap of the correlation between variables



The above Heatmap shows the relationship between the variables of interest in this study. We can see that even though there are some correlations between independent variables and between independent and dependant variables, there are no serious multicollinearity issues. This result was later check with the VIF test that confirmed the observations. Age and income have a positive correlation(0.28) possibly due to higher human capital and experience, income is also positively correlated with certification level(0.18) and negatively to gender(being a

female) (-0.13). between covariables, GPA and institutional reputation were negatively correlated suggesting that more reputable schools might have more rigorous programs. Furthermore we will examine the relationship of these variables in an econometric model.

OLS Model

The primary method used in this paper is a multivariate ordinary least squares(OLS) model that aims to capture the linear relationship between the variables and income of graduates. Our regression model is based on the data from the 2018 and the 2013 NGS. The following are the regression equation and its results:

$$Income=\beta 0 + \beta 1gpa + \beta 2class\ ranking + \beta 3age + \beta 4coop + \beta 5inst\_rep + \beta 6student\_loan + \beta 7gender + \beta 8year\_grad + \beta 9occupation + \beta 10program + \beta 11certification\ level + e$$

Table 3 displays the results of the OLS regression on combined 2018 and 2013 data, with an adjusted R-squared of 16.4%. This suggests that the model explains a significant portion of the income differences among graduates, despite unobservable variables such as abilities and human capital measures. Heteroskedasticity was addressed using the HC3 robust method.

The GPA positively influences income significantly, aligning with literature indicating higher GPAs lead to higher incomes post-graduation. Institution reputation also matters, though its coefficient is lower than GPA's, potentially reflecting a correlation-causation issue.

Student loans have a negative coefficient but lack statistical significance. Non-participation in coop programs negatively impacts outcomes, as expected, although longer graduation times for coop students may slightly depress the coefficient.

Higher certification levels generally lead to higher incomes. Notably, college graduates have a significantly negative coefficient compared to bachelor's or higher degree holders.

The "year_grad" variable indicates lower incomes for 2018 graduates compared to 2013, possibly due to differing economic conditions or skewed samples.

A gender wage gap persists even after controlling for GPA and program studied, warranting further investigation.

Age positively impacts income, reflecting accumulated human capital, education, and work experience among older individuals.

Table 3: OLS model for the Combined cohorts

OLS Regression Results

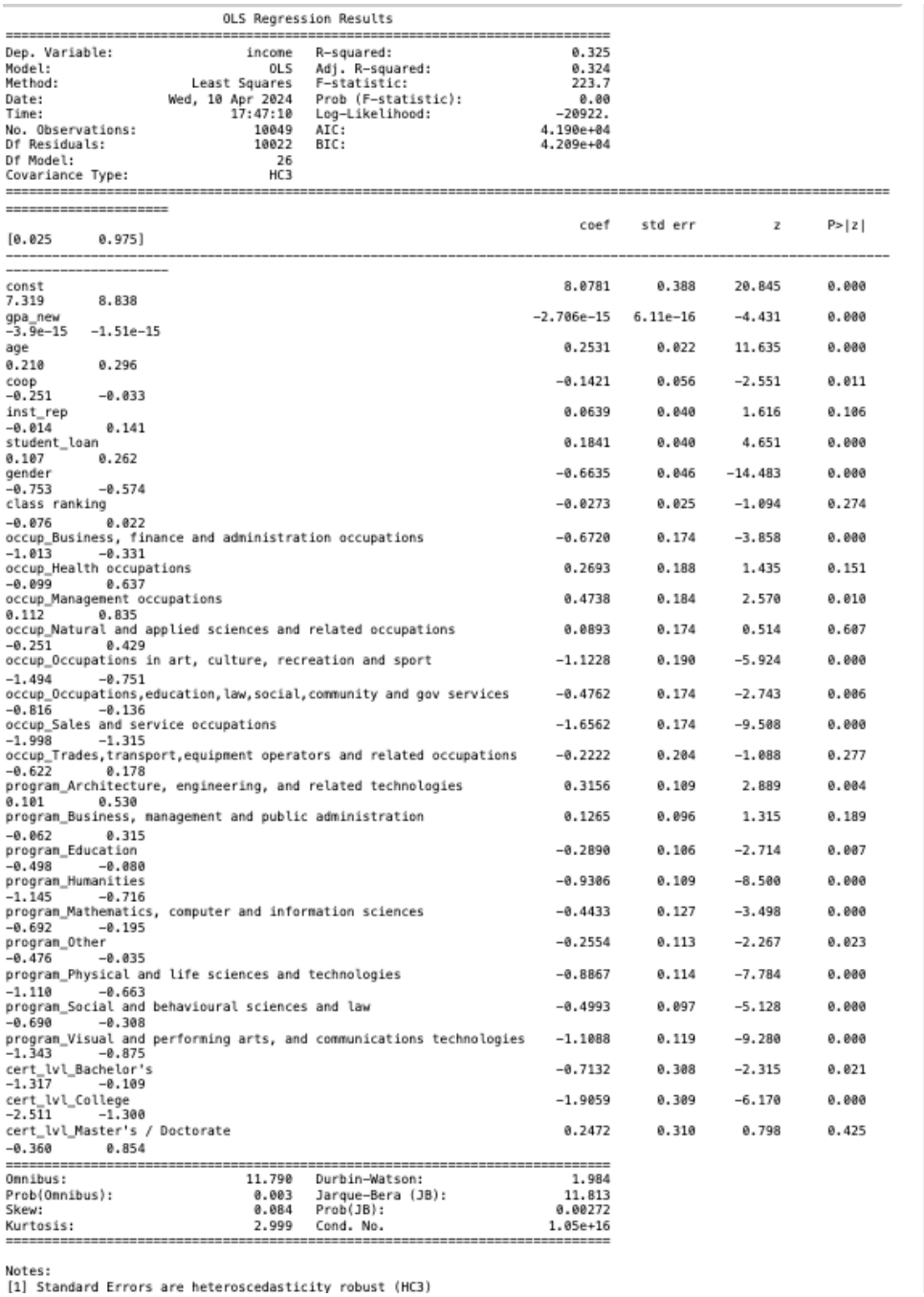
Dep. Variable:	income	R-squared:	0.165
Model:	OLS	Adj. R-squared:	0.164
Method:	Least Squares	F-statistic:	144.8
Date:	Thu, 11 Apr 2024	Prob (F-statistic):	0.00
Time:	17:30:27	Log-Likelihood:	-42884.
No. Observations:	19255	AIC:	8.583e+04
Df Residuals:	19226	BIC:	8.605e+04
Df Model:	28		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
const	6.3977	0.288	22.235	0.000	5.834	6.962
gpa_new	0.5730	0.042	13.496	0.000	0.490	0.656
class ranking	0.0394	0.026	1.520	0.129	-0.011	0.090
age	0.5745	0.017	32.951	0.000	0.540	0.609
coop	-0.1904	0.048	-3.934	0.000	-0.285	-0.096
inst_rep	0.1812	0.035	5.190	0.000	0.113	0.250
student_loan	-0.0697	0.033	-2.139	0.032	-0.134	-0.006
gender	-0.6255	0.035	-18.085	0.000	-0.693	-0.558
year_grad	-1.5165	0.121	-12.550	0.000	-1.753	-1.280
occup_Business, finance and administration occupations	-0.2794	0.126	-2.217	0.027	-0.526	-0.032
occup_Health occupations	0.1902	0.137	1.383	0.167	-0.079	0.460
occup_Management occupations	0.2867	0.135	2.121	0.034	0.022	0.552
occup_Natural and applied sciences and related occupations	0.1076	0.124	0.867	0.386	-0.136	0.351
occup_Occupations in art, culture, recreation and sport	-0.4371	0.143	-3.054	0.002	-0.718	-0.157
occup_Occupations,education,law,social,community and gov services	-0.2342	0.125	-1.873	0.061	-0.479	0.011
occup_Sales and service occupations	-0.7724	0.128	-6.035	0.000	-1.023	-0.522
occup_Trades,transport,equipment operators and related occupations	-0.0008	0.146	-0.006	0.995	-0.287	0.285
program_Architecture, engineering, and related technologies	0.2661	0.088	3.037	0.002	0.094	0.438
program_Business, management and public administration	0.0448	0.078	0.572	0.567	-0.109	0.198
program_Education	-0.1987	0.089	-2.241	0.025	-0.372	-0.025
program_Humanities	-0.6297	0.093	-6.771	0.000	-0.812	-0.447
program_Mathematics, computer and information sciences	-0.1786	0.103	-1.728	0.084	-0.381	0.024
program_Other	-0.0846	0.093	-0.910	0.363	-0.267	0.098
program_Physical and life sciences and technologies	-0.3963	0.091	-4.341	0.000	-0.575	-0.217
program_Social and behavioural sciences and law	-0.2234	0.079	-2.818	0.005	-0.379	-0.068
program_Visual and performing arts, and communications technologies	-0.7230	0.103	-7.042	0.000	-0.924	-0.522
cert_lvl_Bachelor's	-0.0678	0.224	-0.303	0.762	-0.507	0.371
cert_lvl_College	-0.7327	0.225	-3.258	0.001	-1.173	-0.292
cert_lvl_Master's / Doctorate	0.2499	0.225	1.110	0.267	-0.191	0.691

Omnibus:	274.622	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	165.135
Skew:	0.046	Prob(JB):	1.38e-36
Kurtosis:	2.556	Cond. No.	105.

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

OLS based on the 2013 cohort



OLS Regression Results

Dep. Variable:	income	R-squared:	0.325
Model:	OLS	Adj. R-squared:	0.324
Method:	Least Squares	F-statistic:	223.7
Date:	Wed, 10 Apr 2024	Prob (F-statistic):	0.00
Time:	17:47:10	Log-Likelihood:	-20922.
No. Observations:	10049	AIC:	4.190e+04
Df Residuals:	10022	BIC:	4.209e+04
Df Model:	26		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
const	8.0781	0.388	20.845	0.000	7.319	8.838
gpa_new	-2.706e-15	6.11e-16	-4.431	0.000	-3.9e-15	-1.51e-15
age	0.2531	0.022	11.635	0.000	0.210	0.296
coop	-0.1421	0.056	-2.551	0.011	-0.251	-0.033
inst_rep	0.0639	0.040	1.616	0.106	-0.014	0.141
student_loan	0.1841	0.040	4.651	0.000	0.107	0.262
gender	-0.6635	0.046	-14.483	0.000	-0.753	-0.574
class ranking	-0.0273	0.025	-1.094	0.274	-0.076	0.022
occup_Business, finance and administration occupations	-0.6720	0.174	-3.858	0.000	-1.013	-0.331
occup_Health occupations	0.2693	0.188	1.435	0.151	-0.099	0.637
occup_Management occupations	0.4738	0.184	2.570	0.010	0.112	0.835
occup_Natural and applied sciences and related occupations	0.0893	0.174	0.514	0.607	-0.251	0.429
occup_Occupations in art, culture, recreation and sport	-1.1228	0.190	-5.924	0.000	-1.494	-0.751
occup_Occupations,education,law,social,community and gov services	-0.4762	0.174	-2.743	0.006	-0.816	-0.136
occup_Sales and service occupations	-1.6562	0.174	-9.508	0.000	-1.998	-1.315
occup_Trades,transport,equipment operators and related occupations	-0.2222	0.204	-1.088	0.277	-0.622	0.178
program_Architecture, engineering, and related technologies	0.3156	0.109	2.889	0.004	0.101	0.530
program_Business, management and public administration	0.1265	0.096	1.315	0.189	-0.062	0.315
program_Education	-0.2890	0.106	-2.714	0.007	-0.498	-0.080
program_Humanities	-0.9306	0.109	-8.500	0.000	-1.145	-0.716
program_Mathematics, computer and information sciences	-0.4433	0.127	-3.498	0.000	-0.692	-0.195
program_Other	-0.2554	0.113	-2.267	0.023	-0.476	-0.035
program_Physical and life sciences and technologies	-0.8867	0.114	-7.784	0.000	-1.110	-0.663
program_Social and behavioural sciences and law	-0.4993	0.097	-5.128	0.000	-0.690	-0.308
program_Visual and performing arts, and communications technologies	-1.1088	0.119	-9.280	0.000	-1.343	-0.875
cert_lvl_Bachelor's	-0.7132	0.308	-2.315	0.021	-1.317	-0.109
cert_lvl_College	-1.9059	0.309	-6.170	0.000	-2.511	-1.300
cert_lvl_Master's / Doctorate	0.2472	0.310	0.798	0.425	-0.360	0.854

Omnibus:	11.790	Durbin-Watson:	1.984
Prob(Omnibus):	0.003	Jarque-Bera (JB):	11.813
Skew:	0.084	Prob(JB):	0.00272
Kurtosis:	2.999	Cond. No.	1.05e+16

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

To test the generalizability of the data, it was examined on the 2013 dataset by itself. The results showed a very high Adj-R-Squared of 32.4% which means the model explains a large

variation of the data. Of the other significant results I found that management occupations provided the highest paying jobs and the lowest paying occupations were held by those in "sales and services" occupations under 0.1 confidence intervals.

The programs paying the most was "Architecture, engineering and related technologies" programs and the least paying majors are held by "visual arts and performing arts" grads.

Lasso Regression

Lasso Regression Results:
Intercept: 5.453791006804279

	Variable	Coefficient
0	gpa_new	-0.000000
1	age	0.583083
2	coop	-0.000000
3	inst_rep	0.018060
4	student_loan	-0.000000
5	gender	-0.484203
6	occup_Business, finance and administration occ...	-0.000000
7	occup_Health occupations	0.016486
8	occup_Management occupations	0.000000
9	occup_Natural and applied sciences and related...	0.041162
10	occup_Occupations in art, culture, recreation ...	-0.000000
11	occup_Occupations,education,law,social,communi...	-0.000000
12	occup_Sales and service occupations	-0.350842
13	occup_Trades,transport,equipment operators and...	0.000000
14	program_Architecture, engineering, and related...	0.160564
15	program_Business, management and public admini...	0.000000
16	program_Education	-0.000000
17	program_Humanities	-0.000000
18	program_Mathematics, computer and information ...	0.000000
19	program_Other	-0.000000
20	program_Physical and life sciences and technol...	-0.000000
21	program_Social and behavioural sciences and law	-0.000000
22	program_Visual and performing arts, and commun...	-0.000000
23	cert_lvl_Bachelor's	-0.000000
24	cert_lvl_College	-0.428653
25	cert_lvl_Master's / Doctorate	0.175779

To avoid overfitting a lasso regression was used on the model. Since I am using many variables in this study, there's a large chance of overfitting the data. A lasso regression is a regularization

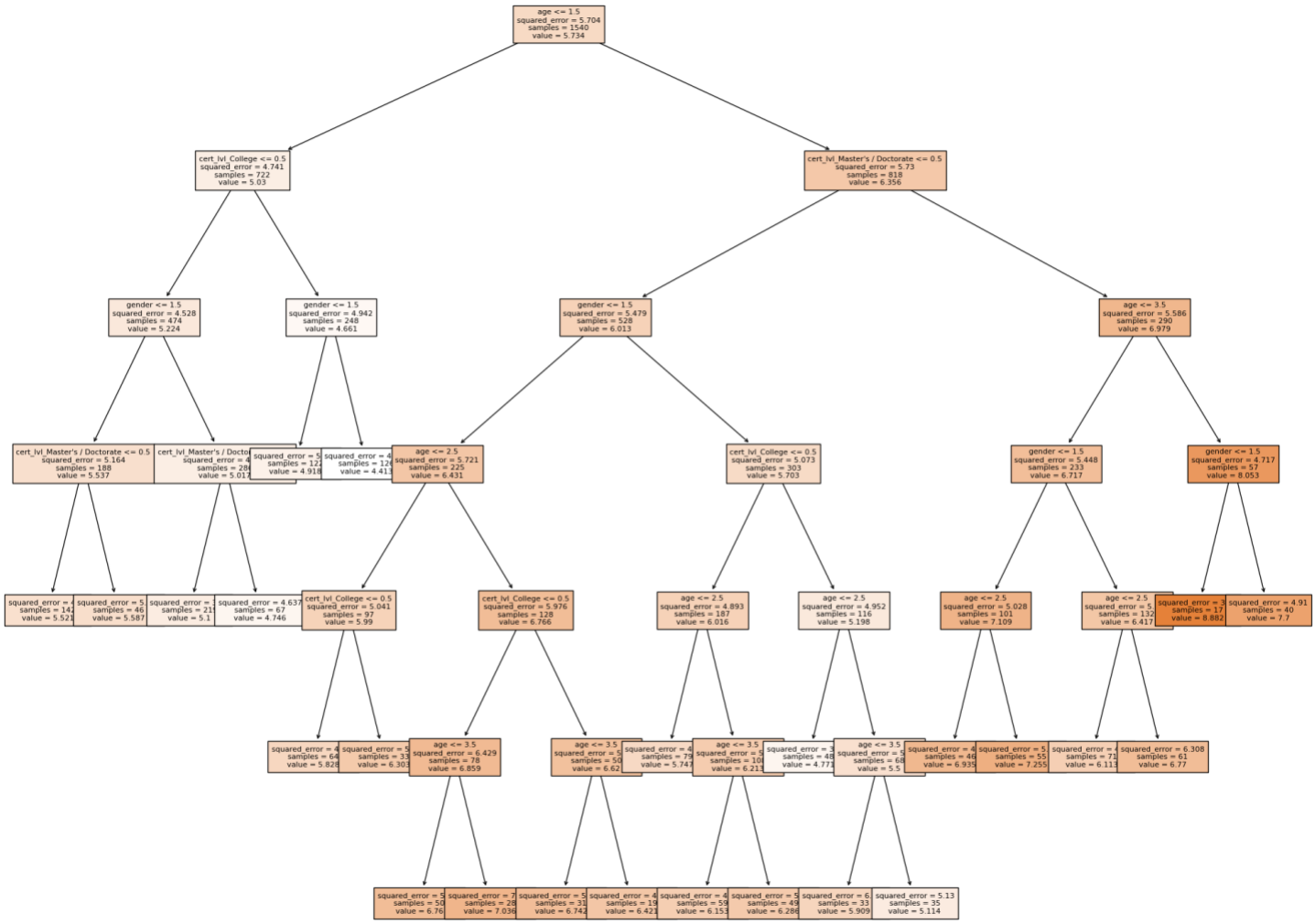
technique in machine learning used for feature selection and preventing overfitting. It adds a penalty term to the standard linear regression cost function, encouraging the model to prefer simpler models by shrinking the coefficients of less important features towards zero.

In this model we find that the lasso regression made the coefficients of many variables in our model zero and the variables remaining are of the highest importance to the predicting of the income of graduates. The final variables remaining are; age, gender, and the certification levels of college and masters as well as a few jobs such as ‘sales and service occupations” and “programs in engineering and architecture”. These variables have the highest amount of impact on the income levels of this sample with respect to other variables.

KNN Regression and Decision trees

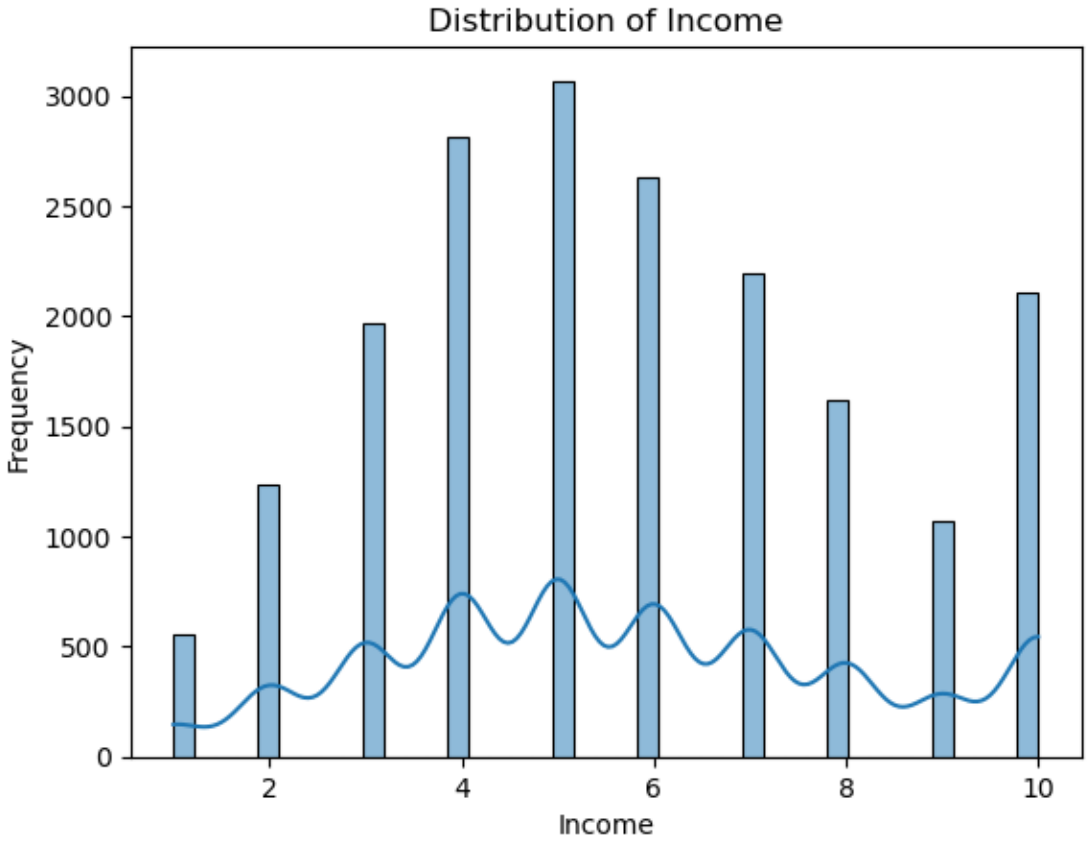
Model	MSE
OLS (Linear Regression)	5.0344244437027899
Decision Tree Regression (Subset)	5.7648095963644 96
KNN Regression (k=1)	9.741893644617 38
KNN Regression (k=10)	5.212957198443 58
KNN Regression (k=20)	5.083891050583658
KNN Regression (k=30)	4.963060959792478
KNN Regression (k=40)	4.973959143968872
KNN Regression (k=50)	5.026794811932556
KNN Regression (k=60)	5.060517725897103
KNN Regression (k=70)	5.077709309404696

Decision Tree Model



Lastly a k-nearest neighbors model was made to train the model on non-parametric models and see which model best describes the data as well as a Decision tree model that used only the variables that had the largest coefficient in from our lasso regression. We found that even though the KNN model with K= 30 had an MSE(4.96), the OLS model(5.03) performed incredibly similar to it and given the fact that the OLS model is simplest to interpret and uses

the least amount of computing power, it is deemed the best way to answer the question at hand and the results stay persistent.



Note: the income levels go up by $10000 intervals- ex. Income level of 6 means the person earns 60000-70000 Cdn dollars per year

Limitations

This paper aims to investigate the factors influencing the income of Canadian graduates, but faces challenges due to time constraints and the encoding of data in ranges instead of actual values, making it difficult for machine learning models and visualizations. While contributing to existing literature, it acknowledges that labor market dynamics are always changing, and new data analysis methods are continually being explored. Despite using variables known to impact early incomes, the results are mostly correlational. The data, drawn from two cohorts of Canadian graduates, aims to be representative but may need adjustments for other populations and cross-validation with additional datasets for generalizability. To predict early career incomes, the paper suggests employing a mix of advanced techniques including various machine learning algorithms like linear regression, decision trees, random forests, gradient boosting machines(GBM), Support vector regression(SVR) and neural networks, along with ensemble methods and hyperparameter tuning like grid search, random search, and Bayesian optimization would be utilized to fine-tune model performance and prevent overfitting and predictive accuracy.

Conclusions

This paper used robust econometric methods found that factors such as higher GPA, higher years of schooling, coop participants, attending reputable institutions and age will positively impact earnings of the graduates. Those with Higher levels of student loans and women tend to have lower earnings post graduation.

Another significant factor was the program that people studied as some program's graduates had a much higher earnings than their counterparts in other programs. This result was same with the field the individual is occupied in as there is a large deviation between the incomes between different occupational paths. The result also found that the year of graduation makes a difference in early career incomes.

References

Card(1999) THE CAUSAL EFFECT OF EDUCATION ON EARNINGS Ch30

Francesconi, Parey(2018) Early Gender gaps among university graduates, European econ.review

Finnie (1999) Early Labour Market Outcomes of Recent Canadian University Graduates by Discipline: A Longitudinal, Cross-Cohort Analysis, Stats Canada Paper No.164

Oreopoulos, Wachter, Heisz(2008) The Short- and Long-Term Career Effects of Graduating in a Recession: Hysteresis and Heterogeneity in the Market for College Graduates, NBER working paper

Dearden, Grewal(2019) Strategic Manipulation of University Rankings, the Prestige Effect, and Student University Choice, NBER working paper

Oosterbeek, Groot, Hartog(1992)AN EMPIRICAL ANALYSIS OF UNIVERSITY CHOICE AND EARNINGS, DE Economist 140

Roksa and Levey(2010)What Can You Do with That Degree? College Major and Occupational Status of College Graduates over Time, Oxford University Press

Rothstein, Rouse(2010) Constrained after college: Student loans and early-career occupational choices

Wyonch(2020)Work-Ready Graduates: The Role of Co-op Programs in Labour Market Success, C.D. Howe Institute

Fitzgerald(2000) College Quality &the Earnings of Recent College Graduates, U.S. Department of Education

Mueller(2020)Grades and Labour Market Earnings in Canada: New Evidence from the 2018 National Graduates Survey, SSRN Electronic Journal

Walters(2004)A Comparison of the Labour Market Outcomes of Postsecondary Graduates of Various Levels and Fields over a Four-Cohort Period, Canadian journal of Sociology

Matkowski(2021) Prediction of Individual Level Income: A Machine Learning Approach

Oehrein (2009)Determining the Future Income of College Students, Wesleyan University

Biewen, Fitzenberger(2006) Employment Effects of Short and Medium Term Further Training Programs in Germany in the Early 2000s, Institut für Arbeitsmarkt- und Berufs- forschung (IAB)

Boissiere, M., Knight, J.B., and Sabot, R.H. "Earnings, Schooling, Ability, and Cognitive Skills." The A merican Economic Review. Dec. 1 985, 75 (5), pp. l016-1030.