**Canadian Wages and Social Change**

Name: Harpreet Kang

Student Number: 500127314

Course: Big Data Analytics Project (CIND820)

Submission Date: May 20, 2024

**Table of Contents**

[Abstract 2](#_Toc167789230)

[Literature Review 3](#_Toc167789231)

[Introduction 4](#_Toc167789232)

[Articles 4](#_Toc167789233)

[1 4](#_Toc167789234)

[2 5](#_Toc167789235)

[3 6](#_Toc167789236)

[Conclusion 7](#_Toc167789237)

[Methodology 7](#_Toc167789238)

[GitHub Repository Link 7](#_Toc167789239)

[Data Understanding 8](#_Toc167789240)

[Data Preparation 8](#_Toc167789241)

[Feature Selection 8](#_Toc167789242)

[Missing Values 8](#_Toc167789243)

[Low Variance 8](#_Toc167789244)

[Exploratory Data Analysis 8](#_Toc167789245)

[Modelling 8](#_Toc167789246)

[Validation 8](#_Toc167789247)

[Approach 10](#_Toc167789248)

[References 10](#_Toc167789249)

# Abstract

According to the 2021 Census, it is estimated that 21% of Canadians earn at least $100, 000 per year (Statista, 2024). In comparison, 9.9% of the population in Canada lives in poverty, with 5% in deep poverty as of 2022 (Government of Canada, Statistics Canada, 2024). With the Consumer Price Index increasing to 2.9% in March 2024 (Government of Canada, Statistics Canada, 2024), many Canadians may experience increased challenges with the cost of living as their spending power declines. With a national average hourly wage of $33.55 in 2023 (Government of Canada, Statistics Canada, 2024), earning power is a conversation of mutual interest to both the government and its labor force. For governments, it drives strategic planning that can create gender equality, lower poverty, and stimulate economic development. For employees, these factors can be used to retrain, transition, or navigate into a second career. Overall, it is imperative to the betterment of both parties that they comprehend these variables. The goal of this paper will be to identify the features most likely to contribute to an hourly wage that is greater than the national average and make suitable policy recommendations for government planning.

This research paper will use the Labour Force Survey: Public Use Micro File (Government of Canada, Statistics Canada, 2024) dataset between January and April 2024 to identify key factors that can be used to predict an employee’s hourly wage. The data is filtered for employees who only have a single full-time job, have a permanent position, and are not in school. These factors can be classified as either human capital, social capital, or labour market (Wang, Liao, & Miu 2022). Limitations of the sample are that the hourly wage is only available for employees – emitting those that are self-employed, and some columns contain missing values. Regarding the latter, feature selection techniques will be first used to reduce the dimensions, such as filtering for low variance, high correlation, and missing values.

Supervised machine learning models such as linear regression and non-linear regression will be the focus of this paper as the dataset contains the target label from which the algorithm can learn. K-Means, a clustering unsupervised machine learning model, will be used to uncover undetected patterns in the dataset for feature engineering. Python will be used as the main programming platform to conduct analyses and computations. Data science packages within Python, such as pandas, numpy, sklearn, ydata\_profiling, and visualization packages, such as matplotlib, will be used to understand and explore the data, conduct predictive modelling, validation, and generate visualizations. Statistical techniques such as parametric and non-parametric tests will be used to identify any significant differences between the machine learning models.

The research questions that will be explored are:

1. What are the key variables that contribute to an hourly wage?
2. Which machine learning model, with tuned hyper parameters, will make the best predictions with a high accuracy rate?
3. Are widely popular categories, such as gender and education, important in determining hourly wage?

# Literature Review

What do you already know about the topic?

What do you have to say critically about what is already known?

Has anyone else ever done anything exactly the same?

Has anyone else done anything that is related?

Where does your work fit in with what has gone before?

Why is your research worth doing in the light of what has already been done?

## Introduction

The freedom to choose your own destiny is the essence of the human spirit. The ability to recognize and change the trajectory of your future based on labour force statistics is a powerful concept. This might be the first study in Canada that uses machine learning to predict a general hourly wage based on data presented in the Labour Force Survey: Public Use Micro File (Government of Canada, Statistics Canada, 2024) (LFS). Thus, it is important to draw insights from similar research, and acknowledge the work of others that could potentially be used in answering the research questions that are present in this paper. Reasoning for selecting certain features is valuable because it gives a path to navigate forward. The 3 types of variables of most important to this research are based on the ideas of human and social capital, and the labour market. Human capital circles around education and training (Goldin 2016), social capital places emphasis on an individual’s social relationships (Putnam 1995), and the labour market ultimately embraces economic cycles.

The variables that this paper will examine are the province of residence, whether an individual resides in a major city, age, sex, marital status, education, private or public sector, immigration status, industry, usual hours worked at the main job, the

## Articles

### 1

In this peer reviewed scholarly study, “Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations” (Matbouli, Y. T., & Alghamdi, S. M. 2022), five machine learning regression techniques were used to predict annual salaries based on many limited surveys from the Saudi labour market, in 2020, that are mapped to the International Standard Classification of Occupations (ISCO-08). The objective was to find the best algorithm for predictions across job titles. A striking difference between this survey and the LFS is that the survey is aggregated and thus the dependent variable, salary, is given as a mean, whereas the LFS provides the hourly wage on an on individual level. Furthermore, the LFS dataset is more robust, as no mappings need to be done with other datasets. The Saudi paper distinguishes between its independent variables as organizational characteristic and occupational features and touches on theories on human capital (merits of an individual) and the labour market (features of an organization). One interesting note on the Saudi paper was that the researchers assumed that high skilled workers have more education and less skilled works have less education, and thus eliminating the education feature. In terms of machine learning, both linear and non-linear regression models were used such as Multiple Linear Regression, Artificial Neural Networks (ANN), Tree Regression, Support Vector Machine Regression, and Bayesian-based machine learning using the Gaussian Process Regression. Also to note that no test of linearity was mentioned, before any machine learning model was used. The performance evaluation metrics used were: root-mean-square error (RMSE), R-squared (R2), and mean absolute error (MAE). The findings of the paper were that non-linear models worked the best such the use of Bayesian ML by applying GPR performed with large and limited training data performed among the best, and ANN performed well when the training data was limited.

### 2

In this second peer reviewed scholarly article, “Prediction of Factors Influencing the Starting Salary of College Graduates Based on Machine Learning” (Wang, Liao, & Miu 2022). looks to uncover the factors that influence salaries of college graduates in China. The data comes from Sichuan, a financial college, between October 2019 and December 31st, 2020. A key discussion in this paper was how certain attributes were classified. The features were categorized either as being associated with human capital, social capital, or the labour market. These groupings allow one more easily to understand how certain factors can have an impact on salary, and to have a broader discussion without worrying about the specificities of a particular attribute. As discussed previously, human capital is discussed in terms of as high academic achievements, social capital is discussed as a social network where resources can be exchanged, and the labour market is discussed as the segmentation between public and private firms. In contrast to the Saudi paper, this paper uses classification techniques to predict whether a starting salary for a college graduate is high or low. The interaction recorded between the three categories was executed in R by five machine learning models: Logistic Regression, Support Vector Machine, Naïve Bayes, Regression Tree, Random Forest, and XGBoost. The latter of this models proves to be the best at determining the starting salary of a college graduate with an accuracy of 92.5%. The metrics for validation are precession, recall, accuracy and F1 score. ROC curve is not used as one of the metrics. The dependent variable was not balanced so sampling method is divided into under sampling and oversampling and combined sampling. From the paper’s correlation analysis, backward stepwise logistic regression, and computing p-values the key findings on what the impacts on salary were: high education levels in fields like engineering, gender, employment characteristics such as in geographical areas, and industries like finance.

### 3

In this third peer reviewed scholarly article, “Machine Learning Based Method for Deciding Internal Value of Talent” . This paper looks at predicting salary using artificial intelligence to help HR making decisions. This paper also recognized that machine learning tolls and salary decision making processes are scarce. The data used in this study was from a European research organization with researchers of all seniorities with 40% holding doctoral degrees. 11 Variables with k-Fold cross validation was used. Different regression models were tested, including linear regression, ridge regression, Lasso regression, SVM, gradient boosting, random forest, neural networks, Bayesian ridge, Ada boost, and KNN. gradient boosting regressor (GBR) wins. All the requirements established in the previous section and the good performance scores led us to select RF as the regression method for recruitment and GBR for salary review. Both methods offer relatively good explainability ([Figure 5](https://journals-scholarsportal-info.ezproxy.lib.torontomu.ca/details/08839514/v36i0001/nfp_mlbmfdivot.xml#F0005) shows the variable importance according to predictors), tend to keep low variance and can compute different input data types (numeric and categorical. Higher education, then experience, publications, projects are important features. Gender is something they didn’t find important. The dataset is very small only 130 references to employees. The goal of the paper was to show the AI can be used to to predict salaries without any biases, and promote decision making consistency

## Conclusion

After completing the literature review of past work, this paper will use many of the same machine learning models to tackle the research questions outlined above. It has become abundantly clear that not much research exists in this topic, especially in Canada, and thus a very naturally compelling argument to conduct this research and present its findings. Surprisingly, not one of these past research papers used a clustering algorithm to detect hidden patterns, and not one used both classification and regression together to compare methodologies. Therefore, both points are of keen interest to explore to tell a comprehensive story.

## Methodology

## GitHub Repository Link

<https://github.com/harpkang/CIND820_CAPSTONE/tree/main>

## Data Understanding

The “Labour Force Survey: Public Use Microdata File” dataset can be found via the Statistic Canada’s website <https://doi.org/10.25318/71M0001X-eng>. The complete dataset that this research is based on is comprised of concatenating 4 files representing 4 months (January, February, March, and April) of 2024. In addition, as an initial starting point to begin the project, any null value within the hourly wage target attribute will be removed to have no missing values. This continuous variable has a precision level of two, is skewed to the right, and is not normally distributed according to the Shapiro-Wilk test. The range of the target variable is between $5.77 and $216.35, and a mean of $34.13. This of course make sense as less of the of the sample will make less than fifty dollars. The refined dataset for this project contains 223, 792 records with 60 columns. We can further reduce this dataset by only focusing on job holders with one job instead of more than 1 which is measured by variable MJH. We can further filter this by only look at non-students using this field SCHOOLN not considering full time or part time students, and then eliminate this variable as well. Filtering one more time for FTPTMAIN to say that we are only looking at full time workers. Lasty we are looking at Permanent positions PERMTEMP=1.Regression can be classified as linear and non-linear. In order to use linear regression, there must be a linear relationship between the independent and dependent variables.

Please see Figure 1 below for the complete list of variables with their respective statistical descriptions. One anomaly that easily stands out is that there are 26 variables that have 60% or more missing values, and 19 of which are missing 100%. These attributes can easily be eliminated from the dataset. In addition, variables record number, survey month, survey year, and standard final weight can also be eliminated because they do not provide any information towards the prediction of target variable. This will leave 30 variables to perform the machine learning models. Marital Status (MARSTAT) is a categorical variable with 6 options. It will be best to binarize this feature as married or not because the data contains 47% married, 30% single, 15% Common-law, with the percentages dwindling downwards. Keeping this variable either 0 or 1 will simplify the algorithm. Labour Force Status (LFSSTAT) variable is also not important because it only contains two categories employed and at work or employed and absent from work. We will assume that every is simply employed regardless of whether one is working or not. The immigrant status variable IMMIG can be further simplified to either immigrant meaning 1 or not 0. We can also remove Tenure, a discrete data type, from the dataset because it is high correlated with Age which is classified into 12 intervals. Variables COWMAIN, prive or public,and UNION are highly correlated and we can delete union. FIrmsize and establishmen size are correlated we can delete the one which is better distributed. Age is something that is universally understood and it can be assumed that as one ages the tenure at the job will be longer aswell.

Three features election techniques that will be used to categorical variables will use the sklearn’s library HotOnesEncoder will be used to transform categorical values into their own feature as either a 0 or 1. NOC\_10 and NOC\_43 are correlated we will pick NOC\_10 as it is more general to pinpoint an industry. The only information we lose is that NOC\_43 breaks down what exactly in the NOC\_10 is. CMA we can binarize into 1 either you are from a big city or not 0.

## Data Preparation

## Feature Selection

### Missing Values

### Low Variance

## Exploratory Data Analysis

## Modelling

## Validation

To use multiple linear regression, there must be an assumption of linearity between the response variable and the independent variables. In our scenario, an anova test can a hypothesis test can be used to check whether the slope of each independent variable is equal to zero or not. If it is equal to zero, then there is no relationship with hourly wage. The equation of the line is:

In addition, four assumptions must be held with random errors . These errors are independent and normally distributed, and have a mean of zero and common variance. This can be checked by

Multiple Linear Regression (MLR) uses the method of least squares to achieve the best fitting line in a hyperplane.

We want to minimize the error when we build a predictive model to make the prediction as accurate as possible.

This is the main goal. Linear regression is used when the variable is numerical discrete or continuous. Least squares to minimize. Relies on assumptions of linearity.

Normalization makes the chance of better preictions. Outliers will reduce our accuracy. Highlyover correlated features will increase the risk of overfitting.

And also, please also check the distribution of the errors in a linear regression.

The distribution of the errors should be normal.

The errors should be normally distributed

When it is a regression problem, logistic regression is used to predict the probabilities of the different possible outcomes of a categorical dependent variable and given a set of independent variables.

so that the initial question about the usefulness of the independent variable x can be restated as:

Is there a linear relationship between x and y?

When multicollinearity is present in a regression problem, it can have these effects on the analysis:

The estimated regression coefficients will have large standard errors, causing imprecision in confidence and prediction intervals.

Adding or deleting a predictor variable may cause significant changes in the values of the other regression coefficients.

How can you tell whether a regression analysis exhibits multicollinearity? Look for these clues:

The value of R2 is large, indicating a good fit, but the individual t-tests are nonsignificant.

The signs of the regression coefficients are contrary to what you would intuitively expect the contributions of those variables to be.

A matrix of correlations, generated by computer, shows you which predictor variables are highly correlated with each other and with the response y.

Supervised learning is the machine-learning task of inferring a function from labelled training data.

So, please check the word "labelled" here. Okay? So, we are inferring a function from the label training data.

Unsupervised learning is the machine-learning task of inferring a function to describe a hidden structure from unlabeled data.

Please here also check the "unlabeled" part, okay?

So, we are trying to understand the, we are trying to describe the hidden structure of the unlabeled data, when we go to unsupervised learning.

Clustering is a descriptive method. We want to use this to explore our data and see patterns in our data. We also mentioned that cluster analysis or clustering is the task of grouping a set of objects in such a way that the objects inside the same group are more similar to each other than to those in other groups.

We will use centroid-based clustering, hard part is selecting k. Looking for hiding structures in the dataset.

## Approach

* When comparing regression you can do an anova test with reduced model-full model.

# References

1. Canada: income distribution (2024 March 11). Statista, Retrieved May 14, 2024, from [https://www.statista.com/statistics/464262/percentage-distribution-of-earnings-in-canada-by-level-of-income/ - :~:text=In 2021, 21.2 percent of, representing the second largest group](https://www.statista.com/statistics/464262/percentage-distribution-of-earnings-in-canada-by-level-of-income/#:~:text=In%202021%2C%2021.2%20percent%20of,representing%20the%20second%20largest%20group)
2. Consumer price index portal (2024 May 10). Government of Canada, Statistics Canada Retrieved May 14, 2024, from <https://www.statcan.gc.ca/en/subjects-start/prices_and_price_indexes/consumer_price_indexes>
3. Dimensions of Poverty Hub (2024 April 26). Government of Canada, Statistics Canada, Retrieved May 16, 2024, from <https://www.statcan.gc.ca/en/topics-start/poverty>.
4. Employee wages by industry, annual (January 2024 05). Government of Canada, Statistics Canada Retrieved, Retrieved May 19, 2024, from <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410006401&pickMembers%5B0%5D=1.1&pickMembers%5B1%5D=2.2&pickMembers%5B2%5D=3.1&pickMembers%5B3%5D=5.1&pickMembers%5B4%5D=6.1&cubeTimeFrame.startYear=2023&cubeTimeFrame.endYear=2023&referencePeriods=20230101%2C20230101>
5. Goldin, Claudia. 2016. "Human Capital." In Handbook of Cliometrics, ed. Claude Diebolt and Michael Haupert, 55-86. Heidelberg, Germany: Springer Verlag. <http://nrs.harvard.edu/urn-3:HUL.InstRepos:34309590>
6. Labour Force Survey: Public Use Micro File. (2024 May 10). Government of Canada, Statistics Canada, Retrieved May 9, 2024, from <https://doi.org/10.25318/71M0001X-eng>
7. Loyarte-López, E., & García-Olaizola, I. (2022). Machine Learning Based Method for Deciding Internal Value of Talent. Applied Artificial Intelligence, 36(1). <https://doi.org/10.1080/08839514.2022.2151160>
8. Matbouli, Y. T., & Alghamdi, S. M. (2022). Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations.Information, 13(10), 495. <https://doi.org/10.3390/info13100495>
9. Pfeffer, J.; Davis-Blake, A. Understanding Organizational Wage Structures: A Resource Dependence Approach. Acad. Manag. J.1987, 30, 437–455. [[Google Scholar](https://scholar.google.com/scholar_lookup?title=Understanding+Organizational+Wage+Structures:+A+Resource+Dependence+Approach&author=Pfeffer,+J.&author=Davis-Blake,+A.&publication_year=1987&journal=Acad.+Manag.+J.&volume=30&pages=437%E2%80%93455)]
10. Putnam, R. D. (1995). Tuning In, Tuning Out: The Strange Disappearance of Social Capital in America. PS: Political Science and Politics, 28(4), 664–683. <https://doi.org/10.2307/420517>
11. Wang, P., Liao, W., Zhao, Z., & Miu, F. (2022). Prediction of Factors Influencing the Starting Salary of College Graduates Based on Machine Learning. Wireless Communications & Mobile Computing (Online), 2022. <https://doi.org/10.1155/2022/7845545>