**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](#_Toc168422637)

[Literature Review 3](#_Toc168422638)

[Introduction 3](#_Toc168422639)

[Articles 4](#_Toc168422640)

[1 4](#_Toc168422641)

[2 5](#_Toc168422642)

[3 6](#_Toc168422643)

[Conclusion 7](#_Toc168422644)

[Methodology 7](#_Toc168422645)

[GitHub Repository Link 7](#_Toc168422646)

[Understanding the Dataset 8](#_Toc168422647)

[Preparing the Dataset 10](#_Toc168422648)

[Exploratory Data Analysis (EDA) 11](#_Toc168422649)

[Feature Selection Techniques 14](#_Toc168422650)

[Outliers 16](#_Toc168422651)

[Random Forest Dimensionality Reduction 18](#_Toc168422652)

[Population Variance 18](#_Toc168422653)

[Modelling 18](#_Toc168422654)

[Normalization 18](#_Toc168422655)

[Validation 19](#_Toc168422656)

[Approach 20](#_Toc168422657)

[References 20](#_Toc168422658)

# 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 original 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 have one single full-time permanent job, and do not attend school. The 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 used to reduce the dimensions, include filtering for low variance, high correlation, and missing values.

Supervised machine learning models such as linear regression, non-linear regression, and knowledge induction 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

## 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) between January and April 2024. 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 (male or female), marital status (true or false), education level, immigration status (true or false), industry, usual hours worked at the main job, the size of the establishment.

## 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 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 workers have less education, and thus eliminating the education feature. In terms of machine learning, both linear and non-linear regression models were used, in the R programming language, 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. The performance evaluation metrics used were: root-mean-square error (RMSE), R-squared (R2), and mean absolute error (MAE). The finding of the paper was that non-linear models worked the best, and ANN was on such model with a coefficient of determination of 94%.

### 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 these 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 was 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” (Loyarte-López & García-Olaizola 2022). This paper looks at predicting salary using artificial intelligence to help HR in making decisions regarding equitable pay based on historical metrics. The dataset used in this research is of roughly a few hundred rows from a European research organization, from January 2021, that has characteristics of its employees, the researchers. Most notably that dataset contains 40% of researchers holding a doctoral degree. The goal of the paper was to identify the key factors that led to the influence of salary. This information would then be used at the time of hiring, and at the time of possible salary increases. The variables in the dataset are all related to personal characteristics such as gender, work, education, and experience. The regression machine learning models that were used were: Linear, Ridge, Lasso, SVM, Gradient Boosting Regressor (GBR), Random Forest, Neural Networks, Bayesian Ridge, Ada Boost, and KNN. In addition, the K-Folds method was utilized with 10 folds for cross-validation to overcome the small sample size. The model that performed the best was the GBR to predict salary increases, and Random Forest was a better model used at the time of recruitment. The evaluation metrics that were used were the coefficient of determination and absolute mean square. Both models mentioned had an R-Squared of greater than 90%. The features that were important were education and professional experience. Interestingly, gender was below a threshold to be considered as important as the variables mentioned. Overall, this paper provides valuable insights in terms of its methodologies used and features it found to be important.

## 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. In addition, common metrics to ass the validity of the models will be used such as R-Squared, ACI, and Root Means Squared. Also, from conducting this literature review it has become abundantly clear that not much research exists in this field, especially in Canada, and thus there is 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. Therefore, given the past research, the methodology below was developed in order to tackle the research questions outlined above.

## Methodology

## GitHub Repository Link

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

## Understanding the Dataset

The “Labour Force Survey: Public Use Microdata File” (LFS) dataset can be found via the Statistic Canada’s website <https://doi.org/10.25318/71M0001X-eng>. The complete dataset that this paper’s research is based on is comprised of concatenating 4 files representing 4 months (January, February, March, and April) of 2024. The unprocessed dataset contains 442, 576 records made up of 60 features. Please see Figure 1 below for a complete list of all the features. The LFS, which is an observational study conducted each month across Canada, is vital as it the basis for determining important economic indicators such as the unemployment rate. This rich dataset is constructed by compiling survey information of individual Canadians each month. The information contains demographic characteristics such as age, gender, employment status, and labour market characteristics such as the employment status, profession, and industry to name a few. The survey is randomly taken of 56, 000 households across Canada of all household members over the age of 15 (Government of Canada, Statistics Canada, 2024). It is mandatory by law, the Statistic Act, to complete the LFS in good faith (Government of Canada, Statistics Canada, 2024). The random and independent sampling design is a combination of stratified sampling, where the provinces are the strata, and then clustering, where geographic areas are further granularized to reduce bias and error (Government of Canada, Statistics Canada, 2024). Stratifying ensures that the diversity of the population is representative, and clustering allows for cost efficiency.

**Figure 1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Data** | **Data Type** |
| AGE\_12 | Five-year age group of respondent | Qualitative | Ordinal |
| AHRSMAIN | Actual hours worked per week at main job | Quantitative | Continuous |
| ATOTHRS | Actual hours worked per week at all jobs | Quantitative | Continuous |
| CMA | Nine largest CMAs | Qualitative | Nominal |
| COWMAIN | Class of worker, main job | Qualitative | Nominal |
| EDUC | Highest educational attainment | Qualitative | Ordinal |
| EFAMTYPE | Type of economic family | Qualitative | Nominal |
| ESTSIZE | Establishment size | Qualitative | Ordinal |
| FINALWT | Standard final weight | Quantitative | Continuous |
| FIRMSIZE | Firm size | Qualitative | Ordinal |
| FTPTMAIN | Full- or part-time status at main or only job | Qualitative | Nominal |
| HRLYEARN | Usual hourly wages | Quantitative | Continuous |
| IMMIG | Immigrant status | Qualitative | Nominal |
| LFSSTAT | Labour force status | Qualitative | Nominal |
| MARSTAT | Marital status of respondent | Qualitative | Nominal |
| MJH | Single or multiple jobholder | Qualitative | Nominal |
| NAICS\_21 | Industry of main job | Qualitative | Nominal |
| NOC\_10 | Occupation at main job (noc\_10) | Qualitative | Nominal |
| NOC\_43 | Occupation at main job (noc\_43) | Qualitative | Nominal |
| PERMTEMP | Job permanency | Qualitative | Nominal |
| PROV | Province | Qualitative | Nominal |
| REC\_NUM | Order of record in file | Qualitative | Nominal |
| SCHOOLN | Current student status | Qualitative | Nominal |
| SEX | Sex of respondent | Qualitative | Nominal |
| SURVMNTH | Survey month | Qualitative | Ordinal |
| SURVYEAR | Survey year | Qualitative | Ordinal |
| TENURE | Job tenure with current employer | Quantitative | Discrete |
| UHRSMAIN | Usual hours worked per week at main job | Quantitative | Continuous |
| UNION | Union status | Qualitative | Nominal |
| UTOTHRS | Usual hours worked per week at all jobs | Quantitative | Continuous |
| HRSAWAY | Hours away from work, part-week absence only | Quantitative | Continuous |
| PAIDOT | Paid overtime hours in reference week | Quantitative | Continuous |
| UNPAIDOT | Unpaid overtime hours in reference week | Quantitative | Continuous |
| XTRAHRS | Number of overtime or extra hours worked | Quantitative | Continuous |
| AGYOWNK | Age of youngest child | Qualitative | Ordinal |
| AGE\_6 | Age in 2 and 3 year groups, 15 to 29 | Qualitative | Ordinal |
| YAWAY | Reason for part-week absence | Qualitative | Nominal |
| PAYAWAY | Paid for time off, full-week absence only | Qualitative | Nominal |
| WKSAWAY | Number of weeks absent from work | Qualitative | Nominal |
| YABSENT | Reason of absence, full week | Qualitative | Nominal |
| AVAILABL | Availability during the reference week | Qualitative | Nominal |
| DURJLESS | Duration of joblessness | Qualitative | Discrete |
| DURUNEMP | Duration of unemployment | Qualitative | Discrete |
| EVERWORK | Identifies if a person has worked in the last year | Qualitative | Nominal |
| FLOWUNEM | Flows into unemployment | Qualitative | Nominal |
| FTPTLAST | Full- or part-time status of last job | Qualitative | Nominal |
| LKANSADS | Unemployed, placed or answered ads | Qualitative | Nominal |
| LKATADS | Unemployed, looked at job ads | Qualitative | Nominal |
| LKEMPLOY | Unemployed, checked with employers directly | Qualitative | Nominal |
| LKOTHERN | Unemployed, other methods | Qualitative | Nominal |
| LKPUBAG | Unemployed, used public employment agency | Qualitative | Nominal |
| LKRELS | Unemployed, checked with friends or relatives | Qualitative | Nominal |
| PREVTEN | Job tenure with previous employer | Quantitative | Discrete |
| PRIORACT | Main activity before started looking for work | Qualitative | Nominal |
| TLOLOOK | Temporary layoff, looked for work during the last four weeks | Qualitative | Nominal |
| UNEMFTPT | Job seekers by type of work sought and temporary layoffs by work status of last job | Qualitative | Nominal |
| WHYLEFTN | Reason for leaving job during previous year (whyleftn) | Qualitative | Nominal |
| WHYLEFTO | Reason for leaving job during previous year (whylefto) | Qualitative | Nominal |
| WHYPT | Reason for part-time work | Qualitative | Nominal |
| YNOLOOK | Reason for not looking for work during the reference week | Qualitative | Nominal |

## Preparing the Dataset

The dataset was filtered by using 4 independent variables and then eliminating them to achieve the desired working dataset. First, the dependent variable ‘HRLYEARN’ was filtered to have all the null values removed. Second, the data set was filtered for individuals that only held one job which was measured by the variable ‘MJH’. Third, the dataset was filtered to only include non-students by using the ‘SCHOOLN’ variable. Fourth, the variable ‘FTPTMAIN’ was used to filter for full-time employees. Finally, the variable ‘PERMTEMP’ was filtered so that the dataset would only have permanent positions. Ultimately, what was achieved was a dataset that represented full-time employees who had one main job, were working full-time on a permanent basis, and were not in school. These actions reduced the number of records in the working dataset to 155, 250. In any type of regression modelling, there is a rule of thumb that states that the number of samples in your dataset should be greater than or equal to 50 + (8 x Independent Variables), which is the case in this paper.

### Exploratory Data Analysis (EDA)

The ydata\_profiling library in python, generates a report that gives a quick synopsis of the features in the dataset with visuals. A summary below was prepared to highlight the descriptive statistics of the 60 features in Figure 2.

**Figure 2**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Median** | **Min** | **Max** | **Standard Deviation** | **Missing Values %** |
| AGE\_12 |  | 6 | 1 | 10 | 2.35 | 0 |
| AHRSMAIN | 362.17 | 400 | 0 | 990 | 134.64 | 0 |
| ATOTHRS | 362.21 | 400 | 0 | 990 | 134.67 | 0 |
| CMA |  |  | 0 | 9 | 2.8 | 0 |
| COWMAIN |  |  | 1 | 2 | 0.45 | 0 |
| EDUC |  | 4 | 0 | 6 | 1.38 | 0 |
| EFAMTYPE |  |  | 1 | 18 | 4.95 | 0 |
| ESTSIZE |  | 2 | 1 | 4 | 1.05 | 0 |
| FINALWT | 315.26 | 220 | 1 | 2795 | 289.2 | 0 |
| FIRMSIZE |  | 4 | 1 | 4 | 1.12 | 0 |
| FTPTMAIN |  |  | 1 | 1 | 0 | 0 |
| HRLYEARN | 37.21 | 32.79 | 5.77 | 208.33 | 18.55 | 0 |
| IMMIG |  |  | 1 | 3 | 0.61 | 0 |
| LFSSTAT |  |  | 1 | 2 | 0.27 | 0 |
| MARSTAT |  |  | 1 | 6 | 2.11 | 0 |
| MJH |  |  | 1 | 1 | 0 | 0 |
| NAICS\_21 |  |  | 1 | 21 | 5.01 | 0 |
| NOC\_10 |  |  | 1 | 10 | 2.75 | 0 |
| NOC\_43 |  |  | 1 | 43 | 13.01 | 0 |
| PERMTEMP |  |  | 1 | 1 | 0 | 0 |
| PROV |  |  | 10 | 59 | 14.36 | 0 |
| REC\_NUM |  |  | 2 | 112082 | 31965.5 | 0 |
| SCHOOLN |  |  | 1 | 1 | 0 | 0 |
| SEX |  |  | 1 | 2 | 0.5 | 0 |
| SURVMNTH |  | 3 | 1 | 4 | 1.12 | 0 |
| SURVYEAR |  | 2024 | 2024 | 2024 | 0 | 0 |
| TENURE | 100.13 | 73 | 1 | 240 | 82.95 | 0 |
| UHRSMAIN | 3.94 | 4 | 3 | 9.9 | 0.58 | 0 |
| UNION |  |  | 1 | 3 | 0.93 | 0 |
| UTOTHRS | 393.8 | 400 | 300 | 990 | 57.9 | 0 |
| HRSAWAY | 15.43 | 0 | 0 | 990 | 48.32 | 7.8 |
| PAIDOT | 8.89 | 0 | 0 | 800 | 36.97 | 7.8 |
| UNPAIDOT | 7.56 | 0 | 0 | 980 | 30.45 | 7.8 |
| XTRAHRS | 16.45 | 0 | 0 | 980 | 47.08 | 7.8 |
| AGYOWNK |  | 2 | 1 | 4 | 1.08 | 56.75 |
| AGE\_6 |  | 5 | 1 | 6 | 1.2 | 84.04 |
| YAWAY |  |  | 0 | 4 | 1.04 | 86.93 |
| PAYAWAY |  |  | 1 | 2 | 0.5 | 92.2 |
| WKSAWAY |  |  | 1 | 99 | 24.74 | 92.2 |
| YABSENT |  |  | 0 | 3 | 0.96 | 92.2 |
| AVAILABL |  |  |  |  |  | 100 |
| DURJLESS |  |  |  |  |  | 100 |
| DURUNEMP |  |  |  |  |  | 100 |
| EVERWORK |  |  |  |  |  | 100 |
| FLOWUNEM |  |  |  |  |  | 100 |
| FTPTLAST |  |  |  |  |  | 100 |
| LKANSADS |  |  |  |  |  | 100 |
| LKATADS |  |  |  |  |  | 100 |
| LKEMPLOY |  |  |  |  |  | 100 |
| LKOTHERN |  |  |  |  |  | 100 |
| LKPUBAG |  |  |  |  |  | 100 |
| LKRELS |  |  |  |  |  | 100 |
| PREVTEN |  |  |  |  |  | 100 |
| PRIORACT |  |  |  |  |  | 100 |
| TLOLOOK |  |  |  |  |  | 100 |
| UNEMFTPT |  |  |  |  |  | 100 |
| WHYLEFTN |  |  |  |  |  | 100 |
| WHYLEFTO |  |  |  |  |  | 100 |
| WHYPT |  |  |  |  |  | 100 |
| YNOLOOK |  |  |  |  |  | 100 |

The response variable is a continuous label that has a precision level of two, is skewed to the right, and is not normally distributed according to the Shapiro-Wilk test. Below in Figure 3 , the relative frequency distribution shows a visual representation of this phenomenon. Furthermore, this distribution resembles an exponential distribution because all the random variables are positive and continuous. The range of the target variable is between $5.77 and $216.35, the median is $32.79, and the mean is $37.21, which is slightly greater than the national average $33.55 (Government of Canada, Statistics Canada, 2024). Outliers and distribution will discussed later in the paper.

**Figure 3**

A graph of a number of wage

Description automatically generated

There are 4 variables that can be binarized to simplify the working dataset. Firstly, the 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 in the remaining classifications. Keeping this variable either 0 or 1 will increase the performance of the algorithm. Secondly, the immigrant status variable (‘IMMIG’) currently has three categories. Categories 1 and 2 describe immigrants who landed less than more than 10 years ago, and the third category is for non-immigrants. Since in this research we are not focusing distinguishing between classes of immigrants, we can again map this feature to either 1 meaning an individual is an immigrant or 0 meaning otherwise. Finally, the variable ‘CMA’, which indicates on the survey whether an individual is from major metropolitan area. The variable has 9 categories that represent major cities in Canada such as Toronto, Vancouver, and Montreal. For efficiency these 9 categories will be marked by 1, if an individual resides in these cities or 0 otherwise.

### Feature Selection Techniques

It is important to preprocess data because the machine learning library sklearn does not working with any missing values, and requires that all attributes be numeric.

#### Missing Data

One anomaly that easily stands out is that there are 20 variables that contain 100% of null values, and 5 variables that contain over 80% of missing values, and 1 variable that contains over 50% of missing data. In total there are 26 variables that can be eliminated from this dataset (the bottom 26 variables in Figure 2 above).

#### Low Information

Furthermore, variables record number, survey month, survey year, and standard final weight (‘REC\_NUM’, ‘SURVMNTH’, ‘SURVYEAR’, ‘FINALWT’) can also be eliminated because they do not provide any information towards the prediction of the target variable. From this analysis, 4 more features can easily be removed from the working dataset. In addition, continuous variables ‘PAIDOT’ and ‘UNPAIDOT’, which represent paid and unpaid overtime hours show low variance because both show at least 79% zero hours. Thus, these two variables can be eliminated from the working dataset as well. The labour force status (‘LFSSTAT’) nominal variable, in the working dataset, contains two categories, which are employed and not absent from work, and employed and absent from work. We can simply assume that everyone is employed, and whether or not an employee is absent or not does not provide any information.

#### Correlation

Multicollinearity in machine learning models, such as regression, can have varying degrees of effects on the outcome of the analysis. The risk of overfitting the model is one of these concerns, which can lead to the model not being able to perform well on unseen data. In addition, the sum of squares for new features will be small because they will contribute little to no information. Thus, it is important to identify and eliminate highly correlated features. Displaying a correlation of 60 variables in matrix visualization is visually unappealing and uninformative, thus table is prepared with the appropriate correlation scores.

The code to generate the profiling report mentioned above has a default value that auto detects data types to apply different correlation methods. Spearman Correlation is used when two numerical features are assessed, and Cramer’s V association coefficient is used for both when two categorical features are assessed, and when numerical and categorical features are assessed together (General settings - YData Profiling n.d.).

There are a few highly correlated features that will be examined. The ordinal characteristic ‘AGE’ classified into 12 categories and the discrete variable ‘TENURE’, which is expressed in months. The nominal variable ‘COWMAIN’, which has two classifications to identify whether the firm is either public or private, and ‘UNION’, which identifies whether a worker is unionized or not. Features ‘COWMAIN’ and ‘NAICS\_21’, which is comprised of twenty-one industry classifications and is a nominal variable. Both ordinal features firm size (‘FIRMSIZE’) and establishment size (‘ESTSIZE’). Characteristics ‘NOC\_10’ and ‘NOC\_43’ are both nominal variables with many categories breaking down industries further and profession and the nominal variable ‘SEX’, which is represents gender as either male or female. Five continuous variables expressed in hours, ‘UTOTHRS’, ‘ATOTHRS’, ‘AHRMAIN’, ‘UHRMAIN’, and ‘XTRAHRS’ that represent overtime, extra, actual, and usual hours worked. Lastly, nominal variables ‘EFAMTYPE’, which is made up of various classifications of family members working or not, and ages of children, and ‘MARSTAT’ meaning marital status. From the working data set nine variables that will be removed due to high correlation are tenure, union, firm size, noc\_43, noc\_10, utothrs, atothrs, ahrmain, and xtrahrs. Figure 4 below shows the remaining variables in the dataset, after the above feature selection techniques.

**Figure 4**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Data** | **Data Type** |
| AGE\_12 | Five-year age group of respondent | Qualitative | Ordinal |
| CMA | Nine largest CMAs | Qualitative | Nominal |
| EDUC | Highest educational attainment | Qualitative | Ordinal |
| ESTSIZE | Establishment size | Qualitative | Ordinal |
| HRLYEARN | Usual hourly wages | Quantitative | Continuous |
| IMMIG | Immigrant status | Qualitative | Nominal |
| MARSTAT | Marital status of respondent | Qualitative | Nominal |
| NAICS\_21 | Industry of main job | Qualitative | Nominal |
| PROV | Province | Qualitative | Nominal |
| SEX | Sex of respondent | Qualitative | Nominal |
| UHRSMAIN | Usual hours worked per week at main job | Quantitative | Continuous |

### Outliers

Since the response variable is not normally distributed, a non-parametric statical technique is needed to first identify and then remove the outliers. The Interquartile Range will be used in this research to remove the outliers in the dataset. It is imperative to remove outliers from the dependent variable because they inevitably reduce the accuracy of the overall model. Figure 4 below shows the box plot for the label of the dataset. The lower bound of the threshold is $0 dollars, and the upper bound of threshold is $79.38. As an hourly wage cannot be negative, the outliers in this research will be considered as greater than the upper threshold mentioned. Out of 155, 250 records it effects only 4, 685, which is less than 5% of the dataset, and thus can be eliminated from the working dataset.

**Figure 5**

A graph with a green and black line

Description automatically generated

**Figure 6**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Median** | **Min** | **Max** | **Standard Deviation** | **Missing Values %** |
| AGE\_12 |  | 6 | 1 | 10 | 2.36 | 0 |
| CMA |  |  | 0 | 1 | 0.48 | 0 |
| EDUC |  | 4 | 0 | 6 | 1.38 | 0 |
| ESTSIZE |  | 2 | 1 | 4 | 1.05 | 0 |
| HRLYEARN | 35.21 | 32 | 5.77 | 79.37 | 14.37 | 0 |
| IMMIG |  |  | 0 | 1 | 0.42 | 0 |
| MARSTAT |  |  | 0 | 1 | 0.5 | 0 |
| NAICS\_21 |  |  | 1 | 21 | 5.02 | 0 |
| PROV |  |  | 10 | 59 | 14.39 | 0 |
| SEX |  |  | 0 | 1 | 0.5 | 0 |
| UHRSMAIN | 3.94 | 4 | 3 | 9.9 | 0.58 | 0 |

### Random Forest Dimensionality Reduction

From the analysis above, we are left with the following variables shown in Figure 5.

**Figure 6**

## Population Variance

**Feature Engineering**

Clustering, on the other hand, is an unsupervised learning task that involves grouping similar data points together based on their inherent patterns or similarities. It aims to discover hidden structures or clusters in the dataset without any prior knowledge or labels.

A real-world example of clustering is customer segmentation for a retail business.

By analyzing customer data such as purchasing behaviour, demographics, and preferences, clustering algorithms can group customers into distinct segments.

This segmentation can help retailers personalize marketing campaigns, target specific customer groups, and optimize product recommendations.

**Clustering**, on the other hand, is used for market segmentation, anomaly detection, document clustering, or recommendation systems, to uncover patterns and structure in the data.

Both techniques have practical significance and applications in a wide range of fields, enabling us to gain insights and make informed decisions from complex datasets.

## Modelling

### Normalization

Normalization is an important topic as it leads to better predictions. Features are on the same scale.

Three features selection 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.

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.

## 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.

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.

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

Talk about minimizing least squares is a common method in deermining the line of best fit.

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.

We use least square to minimize the error.

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. General settings - YData Profiling (n.d.). General settings - YData Profiling, Retrieved June 2, 2024 from <https://docs.profiling.ydata.ai/4.5/advanced_settings/available_settings/#correlations>
6. 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>
7. 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>
8. Labour Force Survey (LFS). (2024 May 31). Government of Canada, Statistics Canada, Retrieved May 31, 2024 from <https://www.statcan.gc.ca/en/survey/household/3701>
9. Labour Force Survey (LFS). (2024 May 31). Government of Canada, Statistics Canada, Retrieved May 31, 2024 from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1541308
10. 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>
11. 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>
12. 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)]
13. 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>
14. 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>
15. <https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3701>