**Predicting the Hourly Wage of Full-Time Employees in Canada**

Name: Harpreet Kang

Student Number: 500127314

Course: Big Data Analytics Project (CIND820)

Submission Date: July 22nd, 2024

**Table of Contents**

# 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 is to identify the most important features that likely contribute to an hourly wage that is greater than the national average.

This research paper will use the data collected by the Labour Force Survey: Public Use Micro File (Government of Canada, Statistics Canada, 2024) 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). To address the anomalies in the dataset, feature selection techniques such as filtering for low variance, high correlation, and missing values were used in dimensionality reduction. These analyses were completed using Python’s data science packages, such as pandas, numpy, sklearn, scipy, xgboost, statsmodels, mlxtend, ydata\_profiling, matplotlib, and seaborn. Furthermore, these libraries help explore data, conduct predictive modelling, validation, and generate visualizations. From a comprehensive investigation, it was determined that the dataset is not linear, and therefore non-linear regression, classification, and knowledge induction models will be used to answer the research questions below. Finally, statistical techniques such as 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. Which classification model, with tuned hyper parameters, will have a high accuracy and f1-score, and balanced precision and recall?
2. What do data mining techniques, specifically association rules, reveal about the link between a high education level and above average hourly earnings?
3. Which non-linear regression model, with tuned hyper parameters, will have a strong RMSE in regard to the continuous response variable?

# 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 following research questions:

1. Which classification model, with tuned hyper parameters, will have a high accuracy and f1-score, and balanced precision and recall?
2. What do data mining techniques, specifically association rules, reveal about the link between a high education level and above average hourly earnings?
3. Which non-linear regression model, with tuned hyper parameters, will have a strong RMSE in regard to the continuous response variable?

The three classes of variables 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 (Matbouli, Y. T., & Alghamdi, S. M. 2022), social capital places emphasis on an individual’s social relationships (Putnam 1995), and the labour market ultimately embraces economic cycles. They will be presented after completing the feature selection techniques.

## Articles

### 1 (Non-Linear Regression)

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 Saudi 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 Regression, and Bayesian-based machine learning using the Gaussian Process Regression. The performance evaluation metrics used were root-mean-square error (RMSE), R-squared () and mean absolute error (MAE). The finding of the paper was that non-linear models worked the best, and ANN was one such model with a of 94%. Overall, this article will help in selecting the non-linear regression models that will be used in this research project.

### 2 (Classification)

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 to more easily 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 (SVM), Naïve Bayes, Regression Tree, Random Forest (RF), and XGBoost. The latter of these models proved to be the best at determining the starting salary of a college graduate with an accuracy of 92.5%. The metrics for validation were precession, recall, accuracy and F1 score. One interesting observation was that the dependent variable was not balanced so the sampling method was divided into under sampling, oversampling and combined sampling. From the paper’s correlation analysis, backward stepwise logistic regression, and p-value calculations, 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. Overall, this article will help in selecting the classification models that will be used in this research project.

### 3 (Linear & Non-Linear Regression)

In this third peer reviewed scholarly article, “Machine Learning Based Method for Deciding Internal Value of Talent” (Loyarte-López & García-Olaizola 2022), the 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. Again, a common theme with the other two articles above is that the independent variables can be categorized into human capital, social capital and the labour market. Regression machine learning models were used in this analysis, and they were: Linear, Ridge, Lasso, SVM, Gradient Boosting Regressor (GBR), RF, Neural Networks, Bayesian Ridge, Ada Boost, and KNN. In addition, K-Folds folds 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 RF 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 winning models had an of greater than 90%. The features that were important were education and professional experience. Interestingly, gender was below a certain threshold to be considered as important. Overall, this paper provides valuable insights in terms of its methodologies used and the features it found to be important.

### 4 (Association Rules)

In this fourth peer reviewed scholarly article, “The influence of computer network technology on national income distribution under the background of social economy” (2021 Zhu and Luo), it uses association rules and k-means to mine the national income dataset, which contains approximately 30, 000 records, from China between 1952 to 2015. The study had ten features, such as salary, education, sex, marital status, occupation, and other variables related to work, to conduct data mining. The support and confidence levels were set at 25% and 80% respectively, and along with the upper antecedent limit set at 5. The paper used two induction algorithms, but the one of interest is the Apriori one. This paper determined that marital status and education had the strongest correlation, respectively, via the association rules.

### 5 (Feature Selection Technique: Correlation)

In this fifth peer reviewed scholarly article, “Reviewing various feature selection techniques in machine learning-based botnet detection” (2024 Baruah, Borah, and Deka V), the topic of interest was correlation when it was selected. The article focused on the use of online bots, their use in illegal enterprises, and their detection by using machine learning models. To fine tune the bot detecting algorithms the article emphasizes on how important feature selection techniques are to building machine learning models. According to the article, feature selection is used for dimensionality reduction so that only relevant features remain. These features can also be further categorized into strongly and weakly relevant, or irrelevant. A strong feature is one that depends solely on the dependent variable. A weak feature is one that is not important, but can become important in some way due to the type of dataset is required by the researcher. At last, an irrelevant feature is one that does not depend on the response variable, but used to create a specific dataset. The article also discusses the difference between redundant and non-redundant features, which depends on correlation. In terms of predicting the hourly wage, in this research paper, it is important to know which features are correlated. If they are correlated, then they are redundant and vice versa. The article goes on discuss many topics outside the scope of this research paper, but will use what is needed to remove redundant features by using correlation.

### 6 (Feature Selection Technique: Missing Data)

In this sixth peer reviewed scholarly article, “Missing data in medical databases: Impute, delete or classify” (2024 Cismondi, Fialho, Reti, Sousa, and Finkelstein), the topic of interest, as it pertains to the research of predicting hourly wage, is how to deal with the missing values. The goal of this peer reviewed article was to develop techniques in dealing with missing values. One interesting fact was that there are essentially only two ways to deal with missing information. One is to impute missing values, and the second is to simply delete all variables with missing information. Another way that this can be put is by stating what information is recoverable and not recoverable. The goal is to preserve as much of the data as possible and to prohibit the loss of any information, otherwise there could serious inadvertent bias. The most striking simple and strategy that was mentioned in the article was that if there were missing values of 50% or greater, the values would highly likely not be missing randomly. Thus, imputing any missing data in this circumstance would appear to be incorrect. Therefore, the model that would be created out of these features would not be as good as a model that eliminated them. There are other details in the article that were discussed such as classifying missing data, but that is beyond what is needed to complete this research paper. Therefore any missing data in a column that amounts to more than 50% will be deleted.

### 7 (Feature Selection Technique: Low Variance)

In this seventh peer reviewed scholarly article, “Feature Selection: A Data Perspective” (2017 Li, Cheng, Wang, Morstatter, Trevino, Tang, and Liu. (2017), a simple principle was mentioned in a paragraph that was relevant to this research paper’s objectives. This guideline was about eliminating features that have low variances. An example that the article gives is that if an attribute has a value of 0 for all the records, it can be eliminated because it cannot help with in determining a class. Overall, the article reviews some of the newer feature selection techniques in recent years and goes into them at depth. Many of the feature selection techniques are beyond the scope of this research paper, and will be simply implementing what was reviewed. Therefore, any column with a low variance will simply be deleted.

## 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 assess the validity of the models will be used in this research will be RMSE for non-linear regression, and accuracy for classification. Majority of the articles identify features such as education, industry, and experience, and sex that may be important in predicting salary. Each article presents an approach to each of the three models that are going to be used in this research paper: non-linear regression, classification, and knowledge induction. When it comes to feature selection techniques, the three simple and powerful approaches that will be used eliminate features later in this research paper will be correlation, missing data, and low variance. In conclusion, from conducting this literature review it has become abundantly clear that not much research exists in this field, especially in Canada, and thus a very naturally compelling argument to conduct this research and present its findings.

# Methodology

Therefore, given the past research, the methodology below was developed to tackle the research questions outlined above.

# GitHub

<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 is comprised of concatenating 4 files representing 4 months (January, February, March, and April) of 2024. The unprocessed multivariate dataset contains 442, 576 records made up of 60 features. Please see Appendix A Figure 1 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.

# 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 is 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. The employee’s main job refers to the job where the usual hours worked far exceeds any other hours worked at one or more other jobs. Furthermore, usual hours refer to what the employee is contractually obligated to work not including overtime. This contrasts with actual hours, which adjusts the usual hours with any absences from work. Working full-time reflects the fact that an employee was working at least thirty hours a week. These steps 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 a 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 ranked by the missing values in Figure 1.

**Figure 1 – Descriptive Summary**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Median** | **Min** | **Max** | **Standard Deviation** | **Missing Values %** |
| AGE\_12 |  | 6 | 1 | 10 |  | 0 |
| AHRSMAIN | 36.22 | 40 | 0 | 99 | 13.46 | 0 |
| ATOTHRS | 36.22 | 40 | 0 | 99 | 13.47 | 0 |
| CMA |  |  | 0 | 9 |  | 0 |
| COWMAIN |  |  | 1 | 2 |  | 0 |
| EDUC |  | 4 | 0 | 6 |  | 0 |
| EFAMTYPE |  |  | 1 | 18 |  | 0 |
| ESTSIZE |  | 2 | 1 | 4 |  | 0 |
| FINALWT | 315.26 | 220 | 1 | 2795 | 289.2 | 0 |
| FIRMSIZE |  | 4 | 1 | 4 |  | 0 |
| FTPTMAIN |  |  | 1 | 1 |  | 0 |
| HRLYEARN | 37.21 | 32.79 | 5.77 | 208.33 | 18.55 | 0 |
| IMMIG |  |  | 1 | 3 |  | 0 |
| LFSSTAT |  |  | 1 | 2 |  | 0 |
| MARSTAT |  |  | 1 | 6 |  | 0 |
| MJH |  |  | 1 | 1 |  | 0 |
| NAICS\_21 |  |  | 1 | 21 |  | 0 |
| NOC\_10 |  |  | 1 | 10 |  | 0 |
| NOC\_43 |  |  | 1 | 43 |  | 0 |
| PERMTEMP |  |  | 1 | 1 |  | 0 |
| PROV |  |  | 10 | 59 |  | 0 |
| REC\_NUM |  |  | 2 | 112082 |  | 0 |
| SCHOOLN |  |  | 1 | 1 |  | 0 |
| SEX |  |  | 1 | 2 |  | 0 |
| SURVMNTH |  | 3 | 1 | 4 |  | 0 |
| SURVYEAR |  | 2024 | 2024 | 2024 |  | 0 |
| TENURE | 100.13 | 73 | 1 | 240 | 82.95 | 0 |
| UHRSMAIN | 39.38 | 40 | 30 | 99 | 5.79 | 0 |
| UNION |  |  | 1 | 3 |  | 0 |
| UTOTHRS | 39.38 | 40 | 30 | 99 | 5.79 | 0 |
| HRSAWAY | 1.54 | 0 | 0 | 99 | 4.83 | 7.8 |
| PAIDOT | 0.89 | 0 | 0 | 80 | 3.7 | 7.8 |
| UNPAIDOT | 0.76 | 0 | 0 | 98 | 3.05 | 7.8 |
| XTRAHRS | 1.64 | 0 | 0 | 98 | 4.71 | 7.8 |
| AGYOWNK |  | 2 | 1 | 4 |  | 56.75 |
| AGE\_6 |  | 5 | 1 | 6 |  | 84.04 |
| YAWAY |  |  | 0 | 4 |  | 86.93 |
| PAYAWAY |  |  | 1 | 2 |  | 92.2 |
| WKSAWAY |  |  | 1 | 99 |  | 92.2 |
| YABSENT |  |  | 0 | 3 |  | 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 |

There are 4 variables that can be binarized to simplify the working dataset. First is the marital status (‘MARSTAT’), which is a categorical variable with six levels. It is best to binarize this feature, as either married or not married, because the working dataset, as is, contains 47% married, 30% single, 15% common-law, with the percentages dwindling downwards in the remaining classifications. Keeping this variable either at 0 or 1 will simplify the algorithm. Secondly, the immigrant status variable (‘IMMIG’) currently has three categories. Categories 1 and 2 describe immigrants who landed less than 10 years ago, and landed more than 10 years ago respectively, and the third category is for non-immigrants. Since this research is not focusing on distinguishing between classes of immigrants, this feature can be mapped 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 a major metropolitan area, 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

### Missing Data

One anomaly that stands out is that there are twenty variables that contain no values at all, five variables that have 80% of its values missing, and one variable that has over 50% of its values missing. In total there are 26 variables that can be eliminated from this dataset (the bottom 26 variables in Figure 1 above). In the sixth article that was reviewed above, this practice is justified because it is a widely held principal to eliminate features that have over 50% of its values missing, and imputing these values would not be correctly justified.

### Low Variance

In article 7 above, the scholarly article mentioned that it was okay to eliminate features that cannot provide information to help in predicting the response variable due to low variance. The variables record number, survey month, survey year, and standard final weight (‘REC\_NUM’, ‘SURVMNTH’, ‘SURVYEAR’, ‘FINALWT’) can be eliminated because they do not provide any information towards the prediction of the target variable. The four variables that were used to filter the dataset, ‘MJH’, ‘SCHOOLN’, ‘FTPTMAIN’, and ‘PERMTEMP’, naturally have an imbalance and will be eliminated from the dataset. These imbalances can be seen in the last four rows of Figure 2 below. In the same figure, the first five variables show a high imbalance and therefore, those variables will be removed from the dataset. In addition, continuous variables ‘PAIDOT’ and ‘UNPAIDOT’, which represent paid and unpaid overtime hours, show a low variance because both have zero hours worked as 79% of their respective distributions. Thus, these two variables can be eliminated from the working dataset. 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. This research paper will assume that everyone is employed, and disregard whether an employee is absent or not because it does not provide any valuable information.

**Figure 2 – Variable Imbalances**

|  |  |
| --- | --- |
| **Variables** | **Imbalance** |
| XTRAHRS | 80% |
| HRSAWAY | 86% |
| PAIDOT | 90% |
| UNPAIDOT | 90% |
| LFSSTAT | 92% |
| FTPTMAIN | 100% |
| MJH | 100% |
| PERMTEMP | 100% |
| SCHOOLN | 100% |

### 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. Thus, it is important to identify and eliminate highly correlated features. This paper will deploy using correlation as a feature selection technique as mentioned in article 5 above to eliminate the redundant features.

Before any correlation analysis was done, the data was split into a training and testing set, 70% and 30% respectfully to avoid in data leakage. The code to generate the profiling report mentioned is through the ydata\_profling module and 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. From analyzing the correlation table generated by the profile report, a correlation warning appears, if the correlation coefficient is greater than the default value of 0.50. This research paper will continue to adhere to this 0.50 default threshold. Please see the correlation matrix below in Figure 3. The matrix was generated, after removing the variables that contained low variance and missing values from above. The heat map makes it easy to identify visually how strong the correlations are between multiple variables at a single glance. Dark blue indicates a strong positive correlation and dark red demonstrates a strong negative correlation.

**Figure 3 – Correlation Matrix**

A diagram of a number of men and women

Description automatically generated with medium confidence

As we can see in Figure 4 below, there are few variables that were highly correlated. All the variables in the left column of table are the variables that were selected to remain in the dataset, and in the right column to be eliminated. The first two ordinal features that are correlated are firm size (‘FIRMSIZE’) and establishment size (‘ESTSIZE’). The difference between the two features is that the variable firm size documents the total number of employees at all locations of the organization, whereas the measure establishment size records the total number of employees at only the individual’s location of employment regardless of whether the employer has other locations. The nominal variables ‘EFAMTYPE’, which is made up of various classifications of family members working or not along with ages of children, and ‘MARSTAT’ classifying marital status are also correlated. The nominal variables ‘COWMAIN’ and ‘NAICS\_21’, are correlated, and their respective classifications identify whether the firm is either public or private, and the North American Industry Classification System. Furthermore, variables ‘NOC\_10’ and ‘NOC\_43’ are nominal variables, with many categories breaking down industries further by profession, that are correlated with the nominal variable ‘SEX’, which represents gender as either male or female. The five other correlated continuous variables expressed in hours were ‘UTOTHRS’, ‘ATOTHRS’, ‘UHRSMAIN’, and ‘AHRMAIN’ represent usual and actual overtime, usual and hours worked at the main job, and any extra hours worked. From utilizing this feature selection technique, the number of variables can be reduced to what is presented in Figure 6 below. The strongest correlations that were observed were between ‘NAICS\_21’ and ‘COWMAIN’, and ‘UHRSMAIN’ and ‘UTOTHRS’.

**Figure 4 – Correlations**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables Kept** | **Variables Removed** | **Correlation** | **Statistical Test Applied** |
| FIRMSIZE | ESTSIZE | 0.515 | Cramer’s V association coefficient |
| MARSTAT | EFAMTYPE | 0.566 | Cramer’s V association coefficient |
| NAICS\_21 | COWMAIN | 0.818 | Cramer’s V association coefficient |
| SEX | NOC\_10 | 0.506 | Cramer’s V association coefficient |
| SEX | NOC\_43 | 0.539 | Cramer’s V association coefficient |
| UHRSMAIN | AHRSMAIN | 0.548 | Spearman Correlation |
| UHRSMAIN | UTOTHRS | 1.00 | Spearman Correlation |
| UHRSMAIN | ATOTHRS | 0.548 | Spearman Correlation |

### Features Selected

Figure 5 shows the twelve features that will be included in all the machine learning models that will be used in this research paper. Since there are only twelve features remaining, there is no further need for dimensionality reduction. A preliminary analysis was conducted using Random Forest’s feature importance method, but there is no justification to eliminate any further features as there are only a few. The graph in Figure 6 is presented to build intuition and domain knowledge to better understand the features and their rankings in accordance with importance. The top 5 features of importance are Tenure, Education, Hours Worked, Sex, and CMA. As a note, both the training and testing sets had the same feature selection techniques, that were mentioned above, applied to them.

**Figure 5 – Selected Features**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Category** | **Unique Values** |
| EDUC | Human Capital | 7 |
| TENURE | Human Capital | - |
| UHRSMAIN | Human Capital | - |
| NAICS\_21 | Labour Market | 21 |
| FIRMSIZE | Labour Market | 4 |
| UNION | Labour Market | 3 |
| SEX | Social Capital | 2 |
| PROV | Social Capital | 10 |
| AGE\_12 | Social Capital | 10 |
| CMA | Social Capital | 2 |
| MARSTAT | Social Capital | 2 |
| IMMIG | Social Capital | 2 |

**Figure 6 – Feature Importance**

A graph showing the number of companies

Description automatically generated with medium confidence

An interesting observation is that the number of classifications of all the categorical variables combined in the working dataset is 63, as shown in Figure 5 above. The number of dummy variables that are needed are , where k is the number of levels for each qualitative variable. Since there are 9 of these variables, then the total amount of extra features in the dataset would amount to 53, ). Adding the two remaining quantitative features, the grand total becomes 55, which means that the total number of combination of subsets that a regression model could be fitted with is .

## Normality

The response variable ‘HRLYEARN’ 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. The first histogram on the left below in Figure 7, shows the relative frequency distribution for this variable and reinforces the findings of the statistical test. Furthermore, this distribution resembles an exponential distribution because all the random variables are positive and continuous. Figure 7 also shows the distributions of the quantitative variables ‘UHRSMAIN’, and ‘TENURE’, the two histograms from the right respectfully. According to the Shapiro-Wilk test, and a quick visual observation, it is safe to conclude that both features are not normally distributed.

**Figure 7 - Histograms**

A graph with blue bars

Description automatically generated

## Assumptions of Linear Regression

Multiple Linear Regression can be presented by the following formula:

, where alpha is the y-intercept, beta is the partial regression coefficient and is the independent variable. In addition, there is a random error term with four assumptions that must be met to use the linear regression models. The assumptions are firstly, that all the error terms must be independent and normally distributed, and secondly, have a common variance and mean equal to zero. The main goal of regression is to minimize the deviations between the predicted and original values, thus optimizing the sum of squares for errors, by using the least squares method. The Durbin-Watson test will be used to check for independence, Shapiro-Wilk test will be used to check for normality, mean of the residuals will be checked to see whether they are equal to zero, and a scatter plot will be used to visualize whether there is a constant variance among the residuals. The research found that the Durbin-Watson test result is 2.01, which means the residuals are independent. The mean of the residuals is 1.71, which indicates that it is very close to 0, thus unbiased. Figure 8 below shows that the residuals are not perfectly normally distributed and shows a slight skewness towards the right. In addition, the Shapiro-Wilk test confirms this non-normality. Finally, Figure 9 below indicates that there is no common variance among the residuals. An attempt at a log transformation of dependent variable did address the normality issue but did not fix the common variance issue. Therefore, linear regression is ruled out, and will keep the log transformed dependent variable.

**Figure 8 – Histogram of Residuals**

A comparison of a graph

Description automatically generated

**Figure 9 – Scatter Plot (Constant Variance)**

A group of blue dots

Description automatically generated with medium confidence

## Log Transformation of the Dependent Variable

The reason for the log transformation is because, as shown above, the residuals are not normally distributed, and the homoscedasticity assumption does not hold. When the dependent variable is log transformed at least the residuals are normally distributed. In addition, working with a normally distributed dependent variable has its advantages such as parametric testing. Furthermore, the transformation will be to the base and to interpret the results, later in the study, it is crucial to remember that the log must be transformed back into its original value. This can easily be done by taking the value of the log to .

## Outliers

The two boxplots in Figure 10 from the left show the outliers highlighted in red, for hourly earnings and usual hours worked respectfully. Interestingly, there are no outliers in the Tenure variable, the right most boxplot in Figure 10. To remove the outliers in each of these features, a non-parametric statical technique is needed to first identify and then remove them. The interquartile range method will be used in this research to remove the outliers in both hourly earnings and usual hours worked variables. It is imperative to remove outliers from the dependent variable because they inevitably reduce the accuracy of the overall model. Out of 108, 675 records from the training set, the number of outliers is 3, 262 and 16, 265, respectfully for both hourly earnings and usual hours worked. The combined outliers are less than 18% of the training dataset, and thus can be eliminated. To understand the effect of removing the outliers from the two variables, the before and after statistics are shown in Figure 11 and 12 respectively.

**Figure 10 - Boxplots**

A graph of a line with numbers and a line

Description automatically generated with medium confidence

**Figure 11 - Before Removing Outliers**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Data** | **Data Type** | **Mean** | **Median** | **Min** | **Max** | **Standard Deviation** |
| HRLYEARN | Quantitative | Continuous | 3.51 | 3.49 | 1.75 | 5.33 | 0.44 |
| UHRSMAIN | Quantitative | Continuous | 39.38 | 40 | 30 | 99 | 5.78 |
| TENURE | Quantitative | Discrete | 99.99 | 72 | 1 | 240 | 82.95 |

**Figure 12 - After Removing Outliers**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Data** | **Data Type** | **Mean** | **Median** | **Min** | **Max** | **Standard Deviation** |
| HRLYEARN | Quantitative | Continuous | 3.53 | 3.51 | 1.93 | 5.33 | 0.44 |
| UHRSMAIN | Quantitative | Continuous | 38.77 | 40 | 35 | 43.5 | 1.87 |
| TENURE | Quantitative | Discrete | 100.89 | 73 | 1 | 240 | 83.02 |

# Approach

## Machine Learning Libraries

Before any machine learning model can be utilized within the sklearn library, it is mandatory that all values be numeric and non-null. Thus, all the qualitative variables must be converted into a 0 or 1. From the sklearn library, the following modules will be used on all the machine learning models presented in this research paper. These are RandomizedSearchCV, KFolds, OneHotEncoder, MinMax Scaler, ColumnTransformer, and Pipeline. RandomizedSearchCV will be used tune hyper parameters, KFolds will be used to measure performance by 5 folds cross validation, OneHotEncoder will be used to convert all the qualitative variables into dummy ones, MinMaxScaler will be used to scale the quantitative features, ColumnTransformer will be used to store the procedures, to initiate OneHotEncoder and MinMaxScaler, and Pipeline will be used to execute the stored procedures along with the machine learning model.

### K-Folds

The reason why a k-fold cross validation will be used is because order is not important as it is in time series. With time series, order is inherent, and the past data could potentially affect future observations. In this research paper, the observations are independent of each other, and do not influence one another. In 5-folds, five of the folds will be trained on, and one-fold will be tested on. This process will be carried out five times, so that each fold has an opportunity to be a test set. At the end of this iteration, there will be five scores, and the average of the scores will be the result. This ensures that any findings are a good representation of the data. In addition, a five-fold cross validation was chosen because it is less computationally expensive than for example ten folds. The added benefit of utilizing five folds for each machine learning model is that the five results of each model can be statistically analyzed using a non-parametric tests, such as the Friedman Test and Kruskal-Wallis Test, to determine whether the results are statistically different from each other.

### Standardization

MinMaxScaler will be used to standardize the numeric independent values as opposed to other methods because this will force the quantitative features to be scaled between of 0 and 1. It is common practice to standardize attributes, before ingesting them into machine learning models. This will ensure that one feature does not dwarf the other in terms of its scale. The two features that are not normally distributed are tenure, and hours worked, as mentioned above. Since these two variables are skewed to the right, it is best to use a non-parametric method to standardize.

### Randomized Grid Search

To increase the performance of the algorithm, the randomize grid search will test a maximum of five random values, thus a maximum of 5 iterations to tune the hyperparameters. The number of folds in each iteration will be five as well. The implementation of Pipelines will help avoid any potential data leakages as the preprocessing steps are replicated identically in each of the training and testing folds of the cross validation.

### Friedman Test

The Friedman test is a non-parametric test to determine whether two or more groups are statistically different or the same. This test will be used for both classification models and non-linear regression models, which will be later discussed in the paper. This test is primarily used, if the errors of the sample are not normally distributed or the errors do not have a common variance centred at zero. As with all non-parametric tests, there is no condition for normality or common variance. The test can also be used, if there are no means to compare, and only single instances of scores are available. The Friedman test follows a chi square distribution with a degrees of freedom equal to k-1. For reference, the formula is presented below.

K is the number of treatments, or the different machine learning models. T is the rank sum. B is the number of blocks, which in this research paper can be one of two items. First when the baseline classification results are examined the metrics accuracy, recall, precision, and f1-score will be considered as the blocks. When the classification models are tuned, then the blocks will be the folds of the five-fold cross validation. Thus, this test will be used twice in comparing the performance of the classification models to determine any statistical differences. In addition, this test will be used for the tuned non-linear regression models when determining whether there is any statistical difference between them. The blocks in the non-linear regression instance will be the folds of the five-fold cross validation.

Furthermore, this test makes one assumption and that is either K or B must be greater than five. In classification there will be six machine learning models tested, assumption K has been met. When analyzing the tuned non-linear regression algorithms, there are 5 machine learning models with 5 blocks each. Although the assumption has not been met, it is very close. In this research paper, the alpha (the possibility of making a false positive prediction) is 5%, and any probability value less than that will mean that the result is statistically significant. Therefore, if the test statistic yields a result less than 5%, then at least two of the models are performing statistically different. The hypothesis test is presented below, and is the same in every instance.

s

### Kruskal-Wallis Test

The Kruskal-Wallis test is a non-parametric test, which requires no assumptions of normality or common variance, to determine whether two or more groups are statistically different or the same. This test will specifically be used for non-linear regression, which will be later discussed in the paper. The test will be used to determine whether there is a statistical significance among the non-linear regression’s initial results. Please note that a parametric test cannot be used because these results, RMSE scores, are single observations for each model and not an aggregation like an average. Thus, only a non-parametric test can be utilized in this scenario. The Kruskal-Wallis test follows a chi square distribution with a degrees of freedom equal to k-1. For reference, the formula is presented below.

N is the number of observations ranked from smallest to largest, and T is the rank sum. Furthermore, this test assumes that the sample sizes are greater than or equal to 5, and that any ties within the ranks are the average rank of the two. Since there are five machine learning models with an RMSE score (n observations), this meets requirements to use the Kruskal-Wallis test. As mentioned in the Friedman Test discussion above, if the test statistic yields a result less than 5%, then at least two of the models are performing statistically different.

## Non-Linear Regression

There is no doubt that supervised machine learning must be implemented because the data contains a quantitative label on which it can be trained and tested on. As determined above, the assumption of linearity fails, and leaves only this type of regression technique to be modelled. The models that will be used are Regression Tree, Random Forest Regressor, Support Vector Regression, KNN, and XGBoost. These models were mentioned above in the literature review in articles 1 and 3. The metric of evaluation will be RMSE as the interpretability is extremely beneficial. It can allow one to understand the deviation of errors from the hourly wage prediction easily because the units will be the same. Thus, the best fitting model would be the one that produces that lowest RMSE. Two other fundamental concepts to remember about the predictions made by the models are that only dependent values, which are within the range of the dataset can be extrapolated and to do so otherwise would be incorrect, and that correlation is not causation.

## Knowledge Induction

This part of the research paper will focus on the Apriori algorithm in Python to uncover any hidden patterns among the features selected above with the response variable. Since the algorithm only works with categorical variables, the two numeric independent variables will need to be transformed along with the dependent variable. All three of these variables will be categorized into five classes each as shown in Figure 13 below. The bins were split using a quantile method to ensure each variable was balanced. Apriori is a rule-based algorithm that is looking for associations between sets of items. This research is interested in data mining what independent variables (‘left hand side’ or antecedent) lead to one of the hourly earning classifications (‘right hand side’ or consequent), . Furthermore, there are three important components to this model that will be discussed. First, ‘support’ measures how many times the LHS variable is present with a particular RHS classification over the entire dataset. Thus, a higher support percentage indicates that there is strong evidence that the variables occur together. Second, ‘confidence’ is the conditional probability of the RHS occurring given that the LHS has occurred. In other words, this is the ‘support’ divided by the occurrence of the LHS in the dataset. Lastly, ‘lift’ is a measure how many times both the LHS and RHS occur together versus how many times the RHS simply occurs on its own in the dataset. As one can imagine, there is a combinatorial explosion even with a small set of variables in any given dataset. To limit the scope of the item sets, two additional principals will be discussed. First, “downward closure” is the minimum support level an item or an itemset needs to continue through the iterative process to be considered as a rule. The iterative process is where first a single item is evaluated against the support threshold, then a combination of two items is evaluated against threshold, and so on until the support threshold cannot be met any further with higher combinatorial spaces. Second, “antimontonicity” is the principal where a subset of items has not met the support threshold, and ultimately that subset cannot continue forward through the iterative process to become a combination of a larger subset. This model will replicate the support, and confidence levels of article four mentioned above.

**Figure 13 – Numeric values classified**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Index** | **Bin Start** | **Bin End** |
| UHRSMAIN | 0 | 1.00 | 21.00 |
| UHRSMAIN | 1 | 21.00 | 53.00 |
| UHRSMAIN | 2 | 53.00 | 106.00 |
| UHRSMAIN | 3 | 106.00 | 200.00 |
| UHRSMAIN | 4 | 200.00 | 240.00 |
| TENURE | 0 | 35.00 | 37.50 |
| TENURE | 1 | 37.50 | 40.00 |
| TENURE | 2 | 40.00 | 43.50 |
| HRLYEARN | 0 | 1.93 | 3.14 |
| HRLYEARN | 1 | 3.14 | 3.38 |
| HRLYEARN | 2 | 3.38 | 3.64 |
| HRLYEARN | 3 | 3.64 | 3.91 |
| HRLYEARN | 4 | 3.91 | 5.33 |

## Classification

Since the linear assumption was violated, another supervised learning technique that will be used is classification. To begin, the response variable ‘HRLYEARN’, which is a continuous variable, needs to be converted into a multi-classification variable. The dependent variable will be converted into 5 bins by using quantiles so that each bin is balanced. This dependent variable can be viewed in Figure 13 above to see the bin sizes. A quick point to mention is that when using the randomized grid search for classification, stratified k-folds will be used by default, instead of the regular k-folds. The same models and metrics that were used in article two, as mentioned in the literature review, will be used in this approach. These models were Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Random Forest (RF), and XGBoost. The metrics were precession, recall, accuracy and f1 score.

# Initial Code

## Classification

### Decision Tree

Information gain is the most important aspect of a decision tree, which is a non-linear classifier. How a tree determines whether a split, using an attribute, gains information is based on the level of impurity before the split, and after the split. The algorithm comes to a halt when there is no more information gain. The two methods for classification to capture information gain are Entropy and the Gini Index. The formulas a presented below.

Entropy is essentially calculating the expected inverse of the log probability of a class in a leaf node. When entropy of a leaf node is equal to zero, when the probability is one, that means the leaf node is pure, which is the goal. In contrast to when the entropy is equal to one, the leaf node has an equal chance of being any class (so not very helpful). Hence, a lower entropy equals a bigger information gain, which is how a feature is selected to be split on.

Gini Index is similar to entropy in a sense that a result of zero means that the leaf node is pure. In both formulas, a probability of one indicates that the leaf node is pure. If a leaf node has equal classes, the Gini impurity will be equal to one-half. Therefore, when the probability of a certain class is high in the leaf node, the lower the Gini impurity will be, and thus larger the information gain.

One thing to remember about a decision tree is that it may overfit the training data. To counter act this, pruning, restricting how deep the tree can grow, is implemented to generalize the algorithm. If the depth of the tree does not affect performance, then perhaps not enough features are being used.

### Random Forest

An ensemble machine learning model, a non-linear classifier, such as the random forest algorithm is a very popular and is made up of many decision trees working together. Overfitting as discussed in decision trees above, is reduced because the number of decision trees utilized. The decision trees within the forest use different rows of data to implement the split criterion, known as bagging, and determine a classification. The majority vote is taken of all the trees, and the classification with the most votes determines the final classification.

### Logistic Regression

This machine learning model is commonly used as the base line model to compare all other classification models against. Before this model can be executed, it is important to remove any outliers and correlated variables. As a point of interest there is no error term in this algorithm, nor is there any check to be done to determine normality. Since in this research there are more than two classes, this should technically be referred to as a multinomial logistic regression. This regression predicts the probability of how likely the dependent variable will occur and classifies according to a threshold. Essentially, logistic regression is a linear model with a logit transform of y. This activation function, , is how a linear equation gets transformed into a probability.

### Support Vector Machine

This type of machine learning model uses decision boundaries on a hyper plane to classify unlabelled data points. The goal of the algorithm is to maximize the distance between a support vector, the nearest training data point to the decision boundary for a category, and all the other support vectors. The number of support vectors in a model is determined by the number of features in the training set. The chasm between the boundary and vector is referred to as the margin. Thus, a greater margin equates to a greater chance of encountering less error.

The hyper parameters C, and radial basis function (rbf) are two important parameters to this model. If the C parameter is large, then the margin becomes small, leading to potential overfitting. Likewise, if the C parameter is small, then the margin becomes large, leading to potential underfitting. A rbf is a kernel, which transforms data into higher dimensions with no limit. Higher dimensions allow the model to be able to separate the data points in a linear fashion.

### Naïve Bayes

Another linear classifier is Naïve Bayes. This model assumes that all features are independent of one another, and it can be explained by the formula below.

It is easy to calculate the probability of Y, because this will be given in the training dataset. This is also referred to as the prior probability. The probability of the features, the evidence, becomes a constant because the assumption is that there is no dependent information. What the algorithm is learning is the probability of the features given the label, the likelihood. The label with the highest probability is how the model determines the classification.

### XGBoost

Another ensemble machine learning model is the XGBoost. This algorithm uses many simple models and combines them to make a single powerful model. It uses gradient boosting, which means it iteratively performs better than the fitted model that came before it, thus boosting the accuracy.

## Non-Linear Regression

### Regression Tree

The split criterion used in this model is variance instead of Gini index or entropy as is in classification mentioned above. Hence, lower the variance in a leaf node, the more homogenous or pure it is.

### Random Forest Regressor

Just like the regression tree, the split criterion used in this model is variance instead of Gini index or entropy. Hence, lower the variance in a leaf node, the more homogenous or pure it is. Again, the classification is determined by a majority vote by all the trees.

### KNN

In this machine learning model, the average distances are taken of all the k neighbours between the training and testing points. The default neighbours that this model uses are five. A very popular method to calculate distance is the Euclidean method given by the formula below.

It is interesting to note that when the number of neighbours is low, the model maybe overfitting. In contrast, when the number of neighbours is high, the model be underfitting. Like all models there must be a balance between over and underfitting.

### Support Vector Regressor

The difference between this model and support vector machine is that this is only used for continuous response variables. The general principal is the same as what was mentioned in the classification version of this model.

### XGBoost

Similar principles surrounding this ensemble machine learning model were mentioned above in the classification version of this algorithm.

## Metrics

### Accuracy

Accuracy is used as a metric for classification models. The metric determines how many predictions out of all the predictions were correct. Correct refers to true positives and true negatives. Accuracy is a good measure to use, if the classes are balanced. Since, a quantile strategy was used to create 5 balanced buckets, this will be a good metric to determine the overall performance of all the classification models. For reference, the formula is shown below.

### Recall

Another metric for classification is recall and is used for determining how many true positives were captured by the model. This metric provides information about whether the model was able to identify all the actual true positives. For reference, the formula is shown below.

### Precision

A third popular metric for classification is precision, which is used to identify out of all the true positives that were actual true positives, how many were predicted accurately. A key point to remember is that there is always a push and pull relationship between precision and recall, which is mentioned above. A model will most likely not be able to have both a high precision and high recall. Therefore, it is important to determine what is important, false positives or false negatives, to see whether which metric to focus on improving. The goal of this research paper will be to focus on false positives, and have them be lower to have more accurate predictions, rather than to focus on false negatives. For reference, the formula is shown below.

### F1-Score

The last classification model that will be discussed for this research paper is the f1-score. The f1-score is a harmonic mean of the two metrics mentioned above, recall and precision. It can be easily observed, by the formula is shown below, that if the either one of these metrics are zero, then the score of the f1 will be zero as well. The metric is popular to use when there is a class imbalance, which is not the case in this research paper. The only way to get a high score, is to have both precision and recall to perform relatively well.

### RMSE (Root Mean Squared Error)

Firstly, it is important to note that for non-linear regression cannot be used as a metric, because the assumption of linearity was violated. Instead for continuous variables, such as hourly earnings, RMSE is a popular metric to use for non-linear regression, and is read as, lower the result the greater the score. One drawback of this metric is that outliers in the dataset can result in large errors, which further supports the analysis above to remove detected outliers. The best part of this metric is that the results are easy to interpret because they are always in the same unit as the label. As mentioned previously, one caveat is that the response variable was log transformed earlier, and therefore any reading of the RMSE computed will need to be exponentiated to determine the true error in dollars. A low score in this research paper would mean that the errors are as close to zero has possible. The formula is presented below for reference.

## Baseline Results

### Classification

Presented below are the baseline results from the classification models that were described above. The model that performed the best in all the metrics discussed earlier was Random Forest classification, as shown in Figure 14 A. A simple Friedman test can be done to determine whether the machine models perform the same or not. The reason a non-parametric test was chosen because none of the samples are normally distributed from a quick Shapiro-Wilk test. The treatments are the six machine learning models, and the blocks are the different metrics. Figure 14 B displays the findings of the Friedman test. From the results of the test, it can be concluded that these models do not perform the same because the probability value is 0.21%, which is highly significant. To answer the first research question, which classification performs the best, it appears that random forest is superior. Although random forest outscores all the other models, it is the second worst performing model, as seen in Figure 14 C below. The worst model when it comes to computational time is support vector, and the best is naïve bayes. Figure 14 D shows the confusion matrix for random forest, and the remaining confusion matrices for the other models can be found in Appendix A Figure 2.

**Figure 14 A- Initial Classification Baseline Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Recall** | **Precision** | **F1-Score** |
| Random Forest | 70.3% | 70.3% | 70.2% | 70.2% |
| Decision Tree | 63.2% | 63.2% | 63.2% | 63.2% |
| XGBoost | 50.2% | 50.2% | 49.4% | 49.6% |
| SVC | 49.0% | 49.0% | 48.3% | 48.5% |
| Logistic Regression | 42.6% | 42.6% | 41.3% | 41.6% |
| MultinomialNB | 39.9% | 39.9% | 38.1% | 38.4% |

**Figure 14 B - Friedman Test for Classification Results**

s

P-Value = 0.21%

**Figure 14 C – Initial Classification Computational Time**

|  |  |  |
| --- | --- | --- |
| **Model** | **Seconds** | **Minutes** |
| SVC | 354.2 | 5.90 |
| Random Forest | 104.9 | 1.75 |
| Decision Tree | 4.3 | 0.07 |
| XGBoost | 2.7 | 0.04 |
| Logistic Regression | 1.9 | 0.03 |
| MultinomialNB | 0.2 | 0.00 |

**Figure 14 D – Confusion Matrix for Random Forest**

A graph with numbers and a bar chart

Description automatically generated with medium confidence

### Non-Linear Regression

Below in Figure 15 A, represent the baseline results from the non-linear regression models that were described earlier. Furthermore, the results of the raw and exponentiated scores appear fairly similar to each other. From Figure 12 above, the standard deviation for the response variable is 0.44. Since all the RMSE scores are less than 0.44, it can be presumed that the models are reasonably accurate. A quick Kruskal-Wallis test in Figure 15 B shows, a probability value of 40.60%, and therefore deem the results not statistically significant. Further analysis such as hyper tuning will need to be done to perform the statistical test again in order to more clearly answer the third research question. Below in Figure 15 C, it shows all the computational times for the non-linear machine learning models. It can be shown that KNN is a clear winner when it comes to performance, and Support Vector is worst performing non-linear regression model.

**Figure 15 A - Initial** **Non-Linear Regression Baseline Results**

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **Exponentiated** |
| Random Forest | 0.26 | $1.30 |
| Support Vector | 0.31 | $1.36 |
| XGBoost | 0.31 | $1.37 |
| KNN | 0.33 | $1.39 |
| Regression Tree | 0.35 | $1.41 |

**Figure 15 B – Kruskal-Wallis Test for Non-Linear Regression Results**

s

P-Value = 40.60%

**Figure 15 C – Initial Non-Linear Regression Computational Time**

|  |  |  |
| --- | --- | --- |
| **Model** | **Seconds** | **Minutes** |
| Support Vector | 210.1 | 3.50 |
| Random Forest | 132.1 | 2.20 |
| Regression Tree | 2.0 | 0.03 |
| XGBoost | 0.3 | 0.00 |
| KNN | 0.1 | 0.00 |

### Knowledge Induction

As mentioned in article four above, the support and confidence levels were intended to be set to 25% and 80% respectively. However, when the support level was set to 25%, the algorithm did not produce any results. The high support did not generate any associations because there are simply no rules that could be met at that threshold. Since the baselines don’t yield any output, the support will be dropped to 5% to generate results. Likewise, the confidence level was set to 1% as 80% was not yielding any results. The reason is that the probability of LHS | RHS is not high enough to require a very high confidence level, and thus no rules were being generated. Lift will be a used a metric to determine how strong the relationship is between the RHS and the LHS. If the value of lift is 1, then there is no association between the antecedent and consequent. If the value of lift is greater than 1, then there is a positive association. If the value of lift is less than 1 and greater than 0, then there is an inverse relationship between the antecedent and consequent. In order to answer the second research question, does a high education, post-secondary or greater, influence above average hourly earnings, which would be bins 3 and 4, lift will be a very important metric in determining positive association. Please see Figure 16 below for the bin sizes. Figure 17 below shows the results generated by the Apriori algorithm, filtering for bins 3 and 4, sorted by the lift column in descending order. Examining Figure 17, it clearly shows that an individual with “Above bachelor’s degree” will be in the highest hourly earnings bin, which is between $49.90 and $206.44. The lift is 2.25 indicating a strong positive association, a confidence of 45.4%, and support of 5.7%. The results are similar to what was found in article 4 above, which is that higher education does indeed increase your hourly earnings. Interestingly, when examining strictly the antecedent as “Post-secondary certificate or diploma”, and the consequent as either the third or fourth hourly earnings bin, the lift is less than 1. This indicates that there is an inverse relationship between the LHS and RHS, and therefore reducing the likelihood of earning an above average wage. From these results, one can conclude that holding a bachelor’s degree is the bare minimum to be earning a wage that is greater than national average of $33.55 in 2023 (Government of Canada, Statistics Canada, 2024).

**Figure 16 – Reproducing Figure 13 Hourly Earning Bins and with the respective $ values**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Response Variable | Index | Bin Start | Bin End | Bin Start in $ | Bin End in $ |
| HRLYEARN | 0 | 1.93 | 3.14 | $6.89 | $23.10 |
| HRLYEARN | 1 | 3.14 | 3.38 | $23.10 | $29.37 |
| HRLYEARN | 2 | 3.38 | 3.64 | $29.37 | $38.09 |
| HRLYEARN | 3 | 3.64 | 3.91 | $38.09 | $49.90 |
| HRLYEARN | 4 | 3.91 | 5.33 | $49.90 | $206.44 |

**Figure 17 – Association Rules Base Line Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Antecedents** | **Consequents** | **Support** | **Confidence** | **Lift** |
| "Above bachelor's degree" | HRLYEARN\_4.0' | 5.7% | 45.4% | 2.25 |
| "Bachelor's degree", 'FIRMSIZE\_More than 500 employees' | HRLYEARN\_4.0' | 5.5% | 34.9% | 1.73 |
| "Bachelor's degree", 'IMMIG\_0' | HRLYEARN\_4.0' | 5.6% | 32.0% | 1.59 |
| "Bachelor's degree", 'UNION\_Non-unionized' | HRLYEARN\_4.0' | 5.3% | 31.8% | 1.58 |
| "Bachelor's degree" | HRLYEARN\_4.0', 'FIRMSIZE\_More than 500 employees' | 5.5% | 21.6% | 1.52 |
| "Bachelor's degree" | HRLYEARN\_4.0', 'UNION\_Non-unionized' | 5.3% | 20.7% | 1.51 |
| "Bachelor's degree" | HRLYEARN\_4.0' | 7.5% | 29.6% | 1.47 |
| "Bachelor's degree", 'IMMIG\_0' | HRLYEARN\_3.0' | 5.1% | 29.0% | 1.46 |
| "Bachelor's degree" | IMMIG\_0', 'HRLYEARN\_4.0' | 5.6% | 22.1% | 1.44 |
| "Bachelor's degree" | HRLYEARN\_3.0' | 6.7% | 26.4% | 1.33 |
| "Bachelor's degree" | IMMIG\_0', 'HRLYEARN\_3.0' | 5.1% | 20.0% | 1.27 |
| Post-secondary certificate or diploma' | HRLYEARN\_3.0', 'CMA\_0' | 5.1% | 13.3% | 1.08 |
| Post-secondary certificate or diploma' | IMMIG\_0', 'HRLYEARN\_3.0' | 6.6% | 17.0% | 1.08 |
| IMMIG\_0', 'Post-secondary certificate or diploma' | HRLYEARN\_3.0' | 6.6% | 20.7% | 1.04 |
| Post-secondary certificate or diploma', 'CMA\_0' | HRLYEARN\_3.0' | 5.1% | 19.5% | 0.98 |
| Post-secondary certificate or diploma' | HRLYEARN\_3.0' | 7.5% | 19.4% | 0.98 |
| Post-secondary certificate or diploma' | HRLYEARN\_4.0' | 5.5% | 14.3% | 0.71 |

# Hyperparameter Tuning

## Classification

After hyperparameter tuning, it is evident that Random Forest is still the superior machine learning model. Below in Figure 18 A, it shows that the accuracy for Random Forest has improved from 70.3% to 73% via utilizing the random grid search. Figure 18 B below shows that the computational time for the Random Forest is only 43 minutes, which is a modest time when compared to the other algorithms, using a 5-fold cross validation. For reference, the confusion matrix for Random Forest is provided in Figure 18 C. Finally, the Friedman test was conducted, as opposed to a parametric test because although the experimental error has a mean of approximately zero and the variation among the different classification models is the same, as shown in Figure 18 D, a quick Shaprio-Wilk test indicated that the residuals do not have a normal distribution. Figure 18 E shows that the p-value of the Friedman test is 0.09%, thus highly significant, meaning that the models do not perform the same. This can clearly be seen in Figure 18 A as the Random Forest model has the highest scores in all the metrics.

**Figure 18 A** – **Final** **Classification Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Recall** | **Precision** | **F1-Score** |
| Random Forest | 73.0% | 73.0% | 72.9% | 72.9% |
| SVC | 53.7% | 53.7% | 53.2% | 53.3% |
| XGBoost | 52.5% | 52.5% | 51.9% | 52.0% |
| Logistic Regression | 42.5% | 42.5% | 41.1% | 41.4% |
| MultinomialNB | 39.4% | 39.4% | 37.7% | 37.9% |
| Decision Tree | 37.9% | 37.9% | 37.1% | 37.4% |

**Figure 18 B – Final Classification Computational Time**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Seconds** | **Minutes** | **Hours** |
| SVC | 22,280.75 | 371.35 | 6.19 |
| Random Forest | 2,610.75 | 43.51 | 0.73 |
| Logistic Regression | 89.87 | 1.50 | 0.02 |
| XGBoost | 71.22 | 1.19 | 0.02 |
| Decision Tree | 10.54 | 0.18 | 0.00 |
| MultinomialNB | 1.98 | 0.03 | 0.00 |

**Figure 18 C – Random Forest Confusion Matrix**

**A blue squares with numbers

Description automatically generated**

**Figure 18 D – Classification Non-Parametric Test**

**A graph with blue dots and a red line

Description automatically generated**

**Figure 18 E – Friedman Test for Tuned Classification Results.**

s

P-Value = 0.09%

## Non-Linear Regression

After hyperparameter tuning, KNN saw the greatest improvement in terms of RMSE. The initial test result was 0.33, but the final result stands at 0.250. Random Forest, which saw a marginal improvement from 0.262, comes in second at a RMSE score of 0.251. Below in Figure 19 A, it shows all the RMSE scores for all the non-linear regression models tuned using a random grid search. Figure 19 B below shows that the computational time while using a 5-fold cross validation for Random Forest was only 5.48 minutes compared to KNN, which was 76 minutes. Although KNN performed slightly better, Random Forest is a clear winner when it comes to efficiency. Like the reasoning presented above to use the non-parametric test such as the Friedman test, the experimental error has a mean of approximately zero, but the variation among the different classification models is not same, as shown in Figure 19 C. Furthermore, a Shaprio-Wilk test indicated that the residuals do not have a normal distribution, which provides support to use a non-parametric test. Figure 19 D shows that the p-value from the Friedman test is 0.37%, meaning that the models do not perform the same. Returning to Figure 19 A, one can conclude that Random Forest is superior because the RMSE scores between KNN and Random Forest are almost identical, but Random Forest is almost 15 times faster as shown in Figure 19 B.

**Figure 19 A** – **Final** **Non-Linear Regression Results**

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **Exponentiated** |
| KNN | 0.250 | 1.28 |
| Random Forest | 0.251 | 1.29 |
| XGBoost | 0.328 | 1.39 |
| Support Vector | 0.333 | 1.40 |
| Regression Tree | 0.359 | 1.43 |

**Figure 19 B – Final Non-Linear Regression Computational Time**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Seconds** | **Minutes** | **Hours** |
| Support Vector | 70,757.7 | 1,179.29 | 19.65 |
| KNN | 4,563.0 | 76.05 | 1.27 |
| Random Forest | 328.7 | 5.48 | 0.09 |
| XGBoost | 75.8 | 1.26 | 0.02 |
| Regression Tree | 6.8 | 0.11 | 0.00 |

**Figure 19 C – Non-Linear Regression Non-Parametric Test**

**A graph with a dotted line

Description automatically generated**

**Figure 19 D – Friedman Test for Tuned Non-Linear Results.**

s

P-Value = 0.37%

## Limitations

As mentioned previously in the report, hyperparameter tuning would be done using RandomizedSearchCV function within sklearn. The computer that is conducting hyperparameter tuning is an Apple Mac Book Air. The specifications of the computer are:

1. 8-Core CPU & GPU
2. 8GB of Unified Memory
3. 16-core Neural engine
4. Apple M2 Chip

The computer cannot process a hundred thousand records, approximately, with a five-fold cross validation technique over five iterations in a modest time. The computational cost is too high, and thus, there is very little benefit to do this type of analysis. In order to conduct better hyperparameter tuning with a balanced cost and benefit approach a more powerful machine is needed for computations. Another limitation was that there weren’t any similar Canadian peer reviewed scholarly articles to glean more insights from.

# Conclusion

Through the analysis presented above in this paper, the research questions were examined in detail and answered. The best machine learning model with hyperparameter tuning for both classification and non-linear regression was the Random Forest. Figure 20 below outlines the best parameters. As one final test, the Random Forest model for both classification and non-linear regression was fit using the training data with the best parameters. It was then used to predict unseen data that was split early during preprocessing. For classification the accuracy score was 72%, and for non-linear regression the RMSE was 0.256. For the question whether higher education leads to a higher wage, it was determined that education, via association rules, is a very important variable when it comes to determining what hourly wage bracket one will fall into.

**Figure 20 – Best Parameters**

|  |  |
| --- | --- |
| **Best Parameters** | |
| **Random Forest Classifier** | **Random Forest Regressor** |
| n\_estimators: 451 | n\_estimators: 451 |
| min\_samples\_split: 3 | min\_samples\_split: 3 |
| min\_samples\_leaf: 1 | min\_samples\_leaf: 1 |
| max\_features: sqrt | max\_features: sqrt |
| max\_depth: 71 | max\_depth: 71 |

# Continuity

Overall, these research questions can be analyzed further. Firstly, more computing power is needed to run more iterations that are required to generate results that use the entire available dataset. Secondly, it would be interesting to use more complex models such as neural networks to see the difference in output compared to classical machine learning models used in this research. Other elements to include in the research would be pairwise interaction between the independent variables, Mathews Correlation, and Briers Index. Regarding the Apriori algorithm, different levels of support and confidence could be tested to find other interesting rules. The work could potentially be improved by utilizing these insights.

# References

1. Baruah S, Borah DJ,  Deka V (2024 February).  Reviewing various feature selection techniques in machine learning-based botnet detection. *Concurrency Computat Pract Exper*.  2024; 36(12):e8076. doi: [10.1002/cpe.8076](https://doi-org.ezproxy.lib.torontomu.ca/10.1002/cpe.8076)
2. 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)
3. 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>
4. 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>.
5. 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>
6. Federico Cismondi, André S. Fialho, Susana M. Vieira, Shane R. Reti, João M.C. Sousa, Stan N. Finkelstein, Missing data in medical databases: Impute, delete or classify?, Artificial Intelligence in Medicine, Volume 58, Issue 1, 2013, Pages 63-72, ISSN 0933-3657, <https://doi.org/10.1016/j.artmed.2013.01.003>.
7. Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. (2017). Feature Selection: A Data Perspective. ACM Comput. Surv. 50, 6, Article 94 (November 2018), 45 pages. <https://doi-org.ezproxy.lib.torontomu.ca/10.1145/3136625>.
8. 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>
9. 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>
10. 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>
11. 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>
12. 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>
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. Xiaoyun Zhu, Shuping Luo, The influence of computer network technology on national income distribution under the background of social economy, Computer Communications, Volume 177, 2021, Pages 166-175, ISSN 0140-3664, <https://doi.org/10.1016/j.comcom.2021.06.025>.

# Appendix A

**Figure 1 – Variable Descriptions (descriptions were collected from the data dictionary)**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 of Residence | 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 |

**Figure 2 – Confusion Matrices for Classification Baseline Results**

# A graph with numbers and a number in a row Description automatically generated with medium confidenceA graph with numbers and a number Description automatically generated with medium confidenceA graph with numbers and a number in a row Description automatically generated with medium confidenceA graph with numbers and a bar chart Description automatically generated with medium confidenceA graph with numbers and a number in blue squares Description automatically generated with medium confidenceA screenshot of a computer screen Description automatically generated

**Figure 2 – Confusion Matrices for Classification after Hyperparameter Tuning**

**A blue squares with numbers

Description automatically generatedA blue squares with numbers

Description automatically generatedA blue squares with numbers

Description automatically generatedA blue squares with numbers and numbers

Description automatically generatedA blue squares with numbers

Description automatically generatedA screenshot of a blue and white color palette

Description automatically generated**