Human Capital Development:

A Trajectory for Socioeconomic Progress and Wellbeing

Predicting Income Level with Machine Learning Algorithms

AN EDX (CYO) PROJECT

SUBMITTED BY

EBOIGBE, UKPONAYE DESMOND

IN PARTIAL FULFILMENT OF A PROFESSIONAL CERTIFICATE IN DATA SCENCE IN THE HARVARDx DATA SCIENCE PROGRAMME

DECEMBER, 2020

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Human capital is an intangible asset or quality not listed on a company's balance sheet. It can be classified as the economic value of a worker's experience and skills. This includes assets like education, training, intelligence, skills, health, and other things employer's value such as loyalty and punctuality. The concept of human capital recognizes that not all labor is equal. But employers can improve the quality of that capital by investing in employees the education, experience, and abilities of employees all have economic value for employers and for the economy as a whole (Kenton and Sonnenshein , 2020).

In a broader sense, human capital is important because it is perceived to increase productivity and thus profitability of a group of people or a country. So the more a country invests in its citizens (i.e., in their education and training, welfare, etc.), the more productive and profitable it could be. A nation or organization is often said to only be as good as its people.

Human Capital and Economic Growth

There is a strong relationship between human capital and economic growth. Because people come with a diverse set of skills and knowledge, human capital can certainly help boost the economy. This relationship can be measured by how much investment goes into people's education.

Some governments recognize that this relationship between human capital and the economy exists, and so they provide higher education at little or no

cost. People who participate in the workforce who have higher education will often have larger salaries, which means they will be able to spend more.

Thus, human capital development is significant for the following reasons:

- Human capital development increases, knowledge, experience and skills of individual at affordable cost.
- Since all labor is not considered equal, employers can improve human capital by investing in the training, education, and benefits of their employees.
- Human capital is an asset that has a significant relationship with economic growth, productivity, and profitability.
- Like any other asset, human capital can depreciate over a period of time if not effectively utilized.

Features that Determine Human Capital

- Skills and qualifications
- Education levels
- Work experience
- Intelligence
- Emotional status (marital, relationship, welfare, etc)
- Personality (hard work, social skills, communication, judgement)
- Creativity (ability to innovate new idea)
- Geography (social peer pressure of local environment can affect expectations and attitudes)

A Brief History of Human Capital

According to (Kenton and Sonnenshein , 2020), the idea of human capital can be traced back to the 18th century. Adam Smith referred to the concept in his book "An Inquiry into the Nature and Causes of the Wealth of Nations," in which he explored the wealth, knowledge, training, talents, and experiences for a nation. Adams suggests that improving human capital through training and education leads to a more profitable enterprise, which adds to the collective wealth of society. According to Smith, that makes it a win for everyone.

In more recent times, the term was used to describe the labor required to produce manufactured goods. But the most modern theory was used by several different economists including Gary Becker and Theodore Schultz, who invented the term in the 1960s to reflect the value of human capacities.

Schultz believed human capital was like any other form of capital to improve the quality and level of production. This would require an investment in the education, training and enhanced benefits of the workforce.

Human Capital Development

Human capital is considered to be one of the most important elements of economic and social progress. Human capital development is the process of improving an organization's employee performance, capabilities and resources. The process of developing human capital requires creating the necessary environments in which the workforce can learn better and apply innovative ideas, acquire new competencies, develop skills, behaviors and attitudes irrespective of their financial background. The tools for creating these opportunities mostly include training, facilitation, coaching and

consulting. The emphasis lies on meeting the needs of workforce and employers.

How to Increase Human Capital

1. **Specialization and Division of Labour:** Specialization allows workers to concentrate on specific tasks and increased specialization of skills. (Though specialization can also lead to boring, repetitive jobs and limited skill development of workers.)
2. **Education:** Basic education to improve literacy and numeracy has an important implication for a basis of human capital.
3. **Vocational Training:** Direct training for skills related to jobs, electrician, plumbing nursing. A skilled profession requires particular vocational training.
4. **A Climate of Creativity:** An education which enables children to think outside the box can increase human capital in a way that 'rote learning' and an impressive accumulation of facts may not.
5. **Infrastructure:** The infrastructure of an economy will influence human capital. Good transport, communication, availability of mobile phones and the internet are very important for the development of human capital in developing economies.
6. **Competitiveness**. An economy dominated by state monopolies is likely to curtail individual creativity and entrepreneurs. An environment which encourages self-employment and the creation of business enables greater use of potential human capital in an economy.

Importance of Human Capital Development

Human Capital is a measure of the skills, education, capacity and attributes of the labour force which influence their productive capacity and earning potential. Human Capital is the knowledge, skills, competencies and other

attributes embodied in individuals or groups of individuals acquired during their life and used to produce goods, services or ideas in market circumstances. For statistical purposes, human capital can be measured in monetary terms as the total potential future earnings of the working age population. (However, this only captures part of human capital and is a limited measure)

Benefits of Human Capital Development

1. **Structural Unemployment**: Individuals whose human capital is inappropriate for modern employers may struggle to gain employment. A major issue in modern economies is that rapid deindustrialization has left many manual workers, struggling to thrive in a very different labour market. Thus the individual might be challenged to learn new skill
2. **Quality of Employment**: In the modern economy, there is increasing divergence between low-skilled, low-paid temporary jobs (gig economy). High-skilled and creative workers have increased opportunities for self-employment or good employment contracts.
3. **Economic Growth and Productivity:** Long-term economic growth depends increasingly on improvements in human capital. Better educated, innovative and creative workforce can help increase labour productivity and economic growth.
4. **Human Capital Flight**: An era of globalization and greater movement of workers has enabled skilled workers to move from low-income countries to higher income countries. This can have adverse effects for developing economies who lose their best human capital.
5. **Resource Management:** Economic growth in countries with limited natural resources, e.g. Japan, Taiwan and South East Asia. Rely on high-

skilled, innovative workforce adding value to raw materials in the manufacturing process.

6. **Sustainability:** This entails the ability to maintain the need for a special or particular skill in the economy. It also concerns what is left for the future generations; whether we leave enough resources, of all kinds, to provide them with the opportunities at least as large as the ones we have had ourselves.

Relationship between Economic Growth and Well-Being

It is assumed that the well-being of a generation (or of a nation) is ultimately a matter of sustaining economic growth. The growth is defined as a long-term expansion in Gross Domestic Product (GDP). Most economists seem to think that economic growth influences well-being. To argue this point, they look at the ranking of countries by level of GDP per-capita and by values of the Human Development Index (which we previously looked at). The rankings look essentially the same: the rank by level of GDP per-capita predicts the ranking by the Human Development Index very precisely. This high correlation between the two rankings seems to suggest that human development is "just" a matter of GDP per-capita and therefore that the well-being of a nation is determined by its rate of economic growth

The critical feature that interacts with economic growth to determine the extent of improvement (or deterioration) in human development and well-being is income inequality. Inequality refers to the difference in wealth or income across individuals in a country. The more unequal the distribution of wealth or income is, the less strong the improvement in human development is for any given rate of economic growth.

This means that if you have two countries whose GDP is expanding at the same rate (say, 4% a year), but inequality is increasing in the first country and decreasing in the second country, then human development will likely improve faster in the second country (Carmignani and Chowdhury 2011).

The Fallacy

Because economic growth has raised living standards around the world, modern economies have lost sight of the fact that the standard metric of economic growth, gross domestic product (GDP), merely measures the size of a nation's economy and doesn't reflect a nation's welfare. Yet policymakers and economists often treat GDP, or GDP per capita in some cases, as an all-encompassing unit to signify a nation's development, combining its economic prosperity and societal well-being. As a result, policies that result in economic growth are seen to be beneficial for society.

The Inclusive Growth

A combination of economic growth and decreasing inequality therefore provides the best possible scenario for the human development and well-being of a country or a generation. To see why, consider that economic growth is the process through which the "pie" gets bigger. Human development requires all individuals to have access to the extra pie. If instead, only a small number of the population can enjoy the extra pie, then their own well-being might increase, but the well-being of the nation as a whole will not (or it will, but to a much smaller proportion).

This is the idea of inclusive growth**:** a significant increase in human development and well-being requires that all individuals (and not just those at the top of income distribution) have access to the benefits of economic

growth. It also entails gender base equity. If that does not happen, and economic growth only benefits a few, then the nation as a whole might end-up being worse off.

Therefore, the challenge to ensure that the next generation is better off is twofold:

1. More economic growth must be generated
2. Inequalities must be reduced.

And neither of these two challenges is a simple one to tackle.

CHAPTER TWO

METHODS AND ANALYSIS

2.1 PROCESS AND WORKFLOW

In this d study, a census data will be used to build a model to predict if the income of any individual in the dataset is greater than or less than or equal to a certain income level per annum (USD 50,000) based on the information available about that individual in the census data.

The dataset used for the analysis is an extraction from the 1994 census data the survey was conducted by Barry Becker and donated to the public site http://archive.ics.uci.edu/ml/datasets/Census+Income. This dataset is popularly called the "Adult" data set. And will be explored in the following order:

1. **Acquire and Read the data:** Downloading the data directly from the source and reading it.
2. **Describe the data:** Specifically the predictor variables (also called independent variables features) from the Census data and the dependent variable which is the level of income (either "greater than USD 50000" or "less than or equal to USD 50000").
3. **Clean the data:** Any data from the real world is always messy and noisy. The data needs to be reshaped in order to aid exploration of the data and modeling to predict the income level.
4. **Explore the independent variables of the data:** A very crucial step before modeling is the exploration of the independent variables. Exploration provides great insights to an analyst on the predicting power of the variable. An analyst looks at the distribution of the variable, how variable it is to predict the income level, what skews it has, etc. In most analytics

project, the analyst goes back to either get more data or better context or clarity from his finding.

5. **Build the prediction model with the training data:** Since data like the Census data can have many weak predictors, for this particular case study I have chosen the non-parametric predicting algorithm of Boosting. Boosting is a classification algorithm (here we classify if an individual's income is "greater than USD 50000" or "less than or equal to USD 50000") that gives the best prediction accuracy for weak predictors. Cross validation, a mechanism to reduce over fitting while modeling, is also used with Boosting.
6. **Validate the prediction model with the testing data:** Here the built model is applied on test data that the model has never seen. This is performed to determine the accuracy of the model in the field when it would be deployed. Since this is a case study, only the crucial steps are retained to keep the content concise and readable.

2.2 DATA ACQUISITION

A temporary file (db) is created and the data was downloaded from the url: `http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data` and stored in the temporary file named "db". The file is read into a variable "raw_data"

2.3 DATA DESCRIPTION

As the training data file does not contain the variable names, the variable names are explicitly specified while reading the data set. While reading the data, extra spaces are stripped. The dataset is read and stored as a data frame of 32561 rows and 15 columns into the variable name: *raw_data*. A high level summary of the data is below.

Summary of the data

```
      Age                Workclass          Final_Weight       Education        Education_num
 Min.   :17.00   Private         :22696   Min.   :  12285   HS-grad     :10501   Min.   : 1.00
 1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827   Some-college: 7291   1st Qu.: 9.00
 Median :37.00   Local-gov       : 2093   Median : 178356   Bachelors   : 5355   Median :10.00
 Mean   :38.58   State-gov       : 1298   Mean   : 189778   Masters     : 1723   Mean   :10.08
 3rd Qu.:48.00   Self-emp-inc    : 1116   3rd Qu.: 237051   Assoc-voc   : 1382   3rd Qu.:12.00
 Max.   :90.00   (Other)         :  981   Max.   :1484705   11th        : 1175   Max.   :16.00
                 NA's            : 1836                     (Other)     : 5134

      Marital_Status                Occupation              Relationship          Race
 Divorced             : 4443   Prof-specialty : 4140   Husband       :13193   Amer-Indian-Eskimo:  311
 Married-AF-spouse    :   23   Craft-repair   : 4099   Not-in-family : 8305   Asian-Pac-Islander: 1039
 Married-civ-spouse   :14976   Exec-managerial: 4066   Other-relative:  981   Black             : 3124
 Married-spouse-absent:  418   Adm-clerical   : 3770   Own-child     : 5068   Other             :  271
 Never-married        :10683   Sales          : 3650   Unmarried     : 3446   White             :27816
 Separated            : 1025   (Other)        :10993   Wife          : 1568
 Widowed              :  993   NA's           : 1843

      Sex            Capital_Gain    Capital_Loss     Hours_Per_week  Native_Country        Income_Level
 Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00   United-States:29170   <=50K:24720
 Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00   Mexico       :  643   >50K : 7841
                Median :    0   Median :   0.0   Median :40.00   Philippines  :  198
                Mean   : 1078   Mean   :  87.3   Mean   :40.44   Germany      :  137
                3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00   Canada       :  121
                Max.   :99999   Max.   :4356.0   Max.   :99.00   (Other)      : 1709
                                                                 NA's         :  583
```

>

Variable classes and data types.

```
'data.frame':	32561 obs. of  15 variables:
 $ Age           : int  39 50 38 53 28 37 49 52 31 42 ...
 $ Workclass     : Factor w/ 8 levels "Federal-gov",..: 7 6 4 4 4 4 4 6 4 4 ...
 $ Final_Weight  : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
 $ Education     : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...
 $ Education_num : int  13 13 9 7 13 14 5 9 14 13 ...
 $ Marital_Status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
 $ Occupation    : Factor w/ 14 levels "Adm-clerical",..: 1 4 6 6 10 4 8 4 10 4 ...
 $ Relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
 $ Race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
 $ Sex           : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...
 $ Capital_Gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
 $ Capital_Loss  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Hours_Per_week: int  40 13 40 40 40 40 16 45 50 40 ...
 $ Native_Country: Factor w/ 41 levels "Cambodia","Canada",..: 39 39 39 39 5 39 23 39 39 39 ...
 $ Income_Level  : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...
```

Dependent Variable

The dependent variable is “Income Level”, representing the level of income. A value of “<=50K” indicates “less than or equal to USD 50,000” annual earning and “>50K” indicates “greater than USD 50,000” annual earning.

Independent Variable

Below are the independent variables (features or predictors) from the Census Data

Variable Name	Description	Type	Possible Values
Age	Age of the individual	Continuous	Numeric
Workclass	Class of Work	Categorical	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwgt	Final Weight Determined by Census Org	Continuous	Numeric
Education	Education of the individual	Ordered Factor	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Education-num	Number of years of education	Continuous	Numeric
Marital-status	Marital status of the individual	Categorical	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Occupation of the individual	Categorical	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Relationship	Present relationship	Categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried

Variable Name	Description	Type	Possible Values
Race	Race of the individual	Categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Sex of the individual	Categorical	Female, Male
Capital-gain	Capital gain made by the individual	Continuous	Numeric
Capital-loss	Capital loss made by the individual	Continuous	Numeric
Hours-per-week	Average number of hours spent by the individual on work	Continuous	Numeric
Native-country	Average number of hours spent by the individual on work	Categorical	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

2.4 CLEANING THE DATA

The training data set is cleaned for missing or invalid data: About 7.4% (2399/32561) of the dataset has NAs in them. It is observed that most of the missing data occurred in ‘Workclass’ and ‘Occupation’ variables. And the remaining have ‘Native_Country’ variable missing. This could be handle by

imputing data in missing values but 'Workclass', 'Occupation' and 'Native_Country' could potentially be very good predictors of income, imputing data may simply skew the model result, hence the rows with missing values are excluded. The cleaned data (without NA's) is assigned variable name: *new_data*

Also, since most of the missing data 2066/2399 (~86%) rows pertain to the "<=50K" Income_Level and the dataset is predominantly of "<=50K" Income_Level, there will not be much information loss for the predictive model building if data sets with NAs are removed.

2.5 DATA EXPLORATION

Each of the variables will be explored for coincidences, distribution, variance, and predictability. Due to data type because most of the data are non-parametric, transformation of variables would not be necessary to address skewness, instead this study will try to understand the data to determine each variable's predictability. In this section, six categories of data: age, gender, education level, workclass, occupation and work duration will be examined because of their significance in human capital development.

To begin data exploration and creating machine learning algorithms, there is need to partition the dataset into training and test datasets for model building and testing. Thus fifty percent (`p = 0.5`) of the dataset will be extracted for model training while the remainder will be used for testing. The choice model in this study would be *Boosting*, which is non-parametric and does not follow any statistical distribution. The goal would be to predict income level using other demographic parameters as presented in the data set.

2.5.1 The training data (head)

	Age	Workclass	Final_Weight	Education	Education_num	Marital_Status	Occupation	Relationship
2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband
11	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband
14	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family
16	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband
17	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child
18	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried

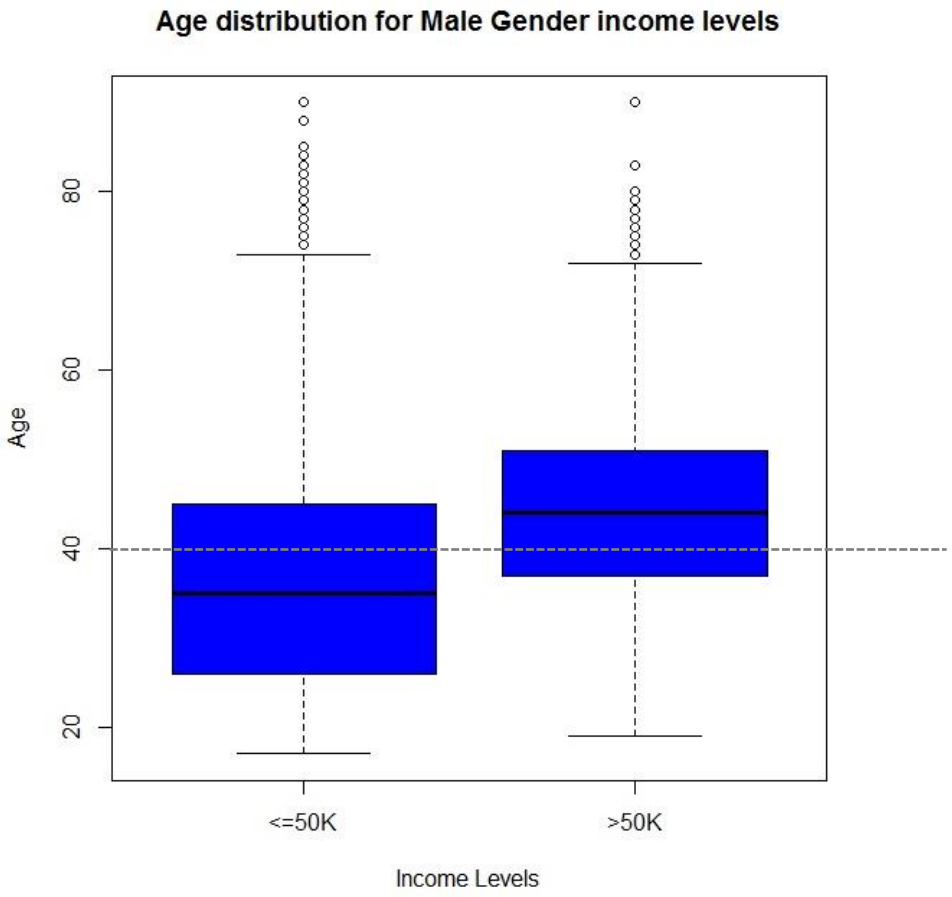
	Race	Sex	Capital_Gain	Capital_Loss	Hours_Per_week	Native_Country	Income_Level
2	White	Male	0	0	13	United-States	<=50K
11	Black	Male	0	0	80	United-States	>50K
14	Black	Male	0	0	50	United-States	<=50K
16	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
17	White	Male	0	0	35	United-States	<=50K
18	White	Male	0	0	40	United-States	<=50K

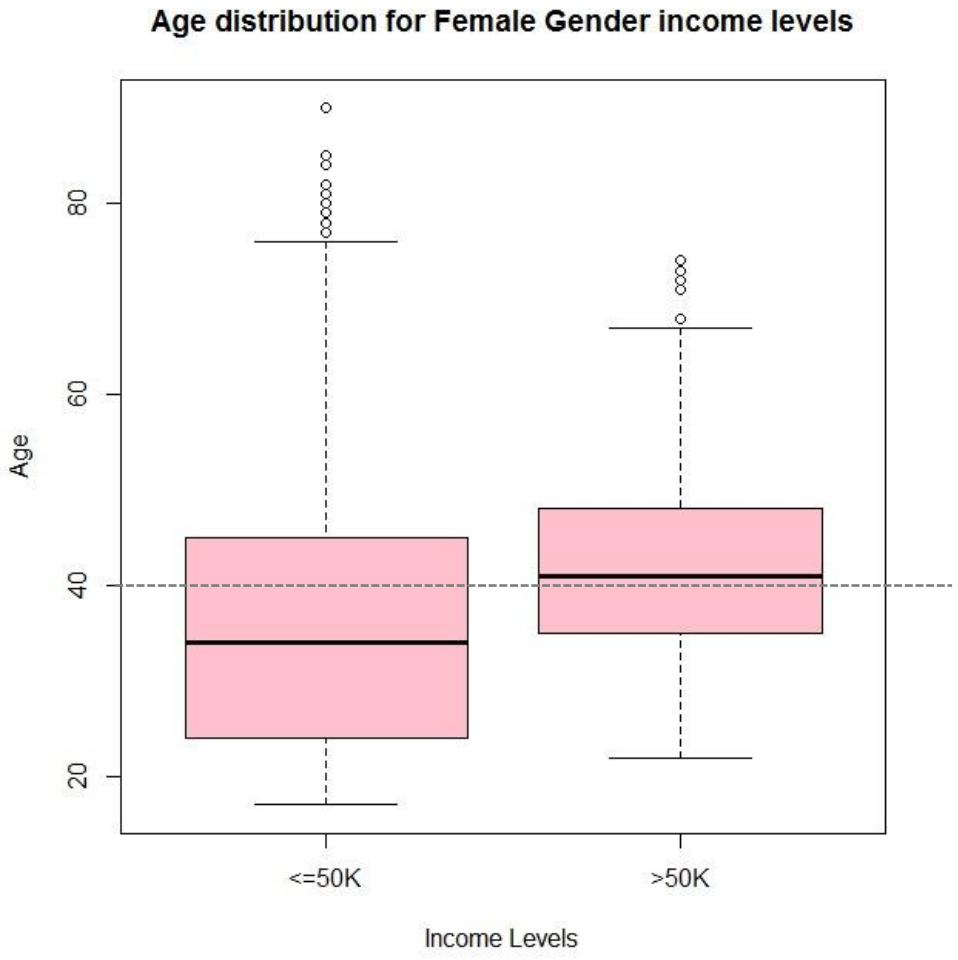
2.5.2 Exploring data by gender

To show the gender proportion as regards income level see table below:

	Income_Level	
Sex	<=50K	>50K
Female	4391	555
Male	6936	3199

It is obvious that the gender proportion in the dataset is `2.049131`, that is, there are slightly above twice males as much as females.

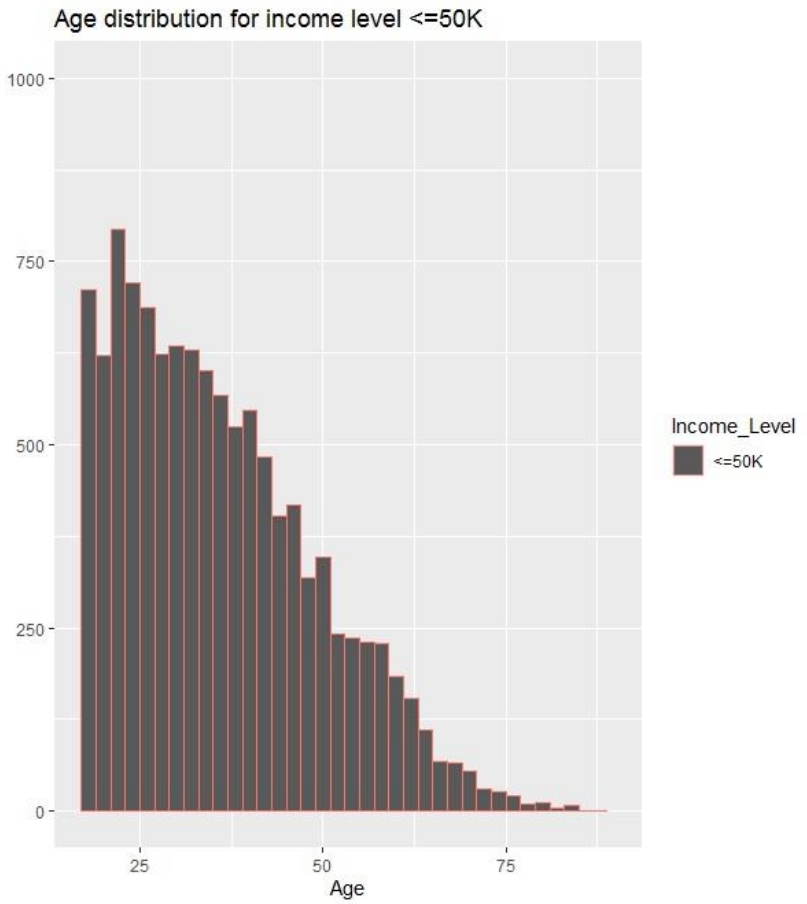


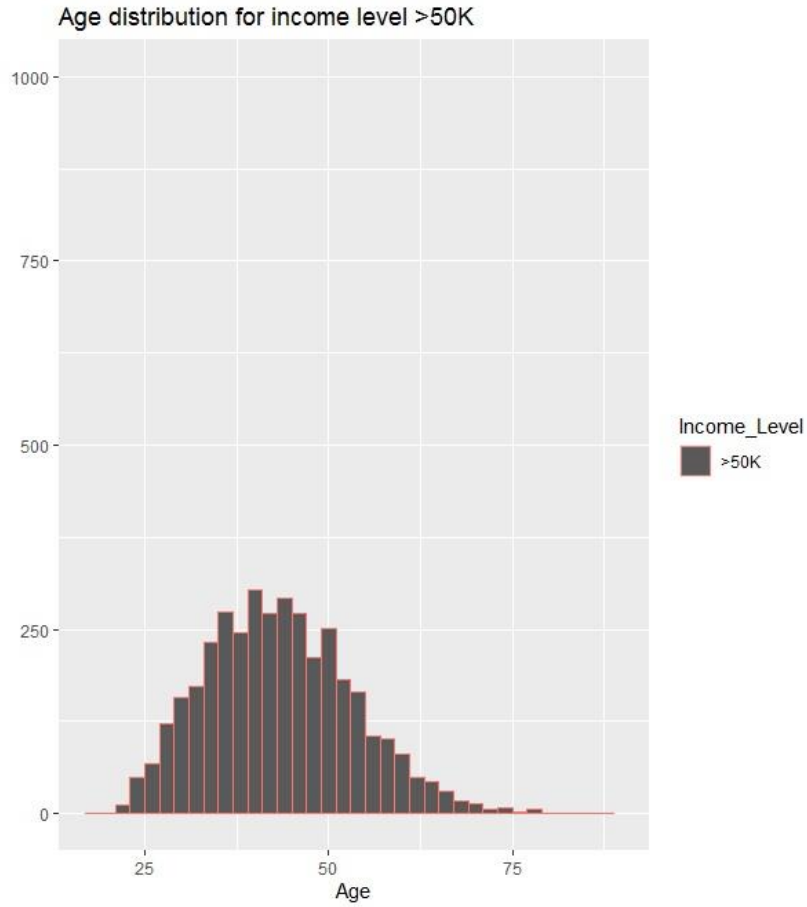


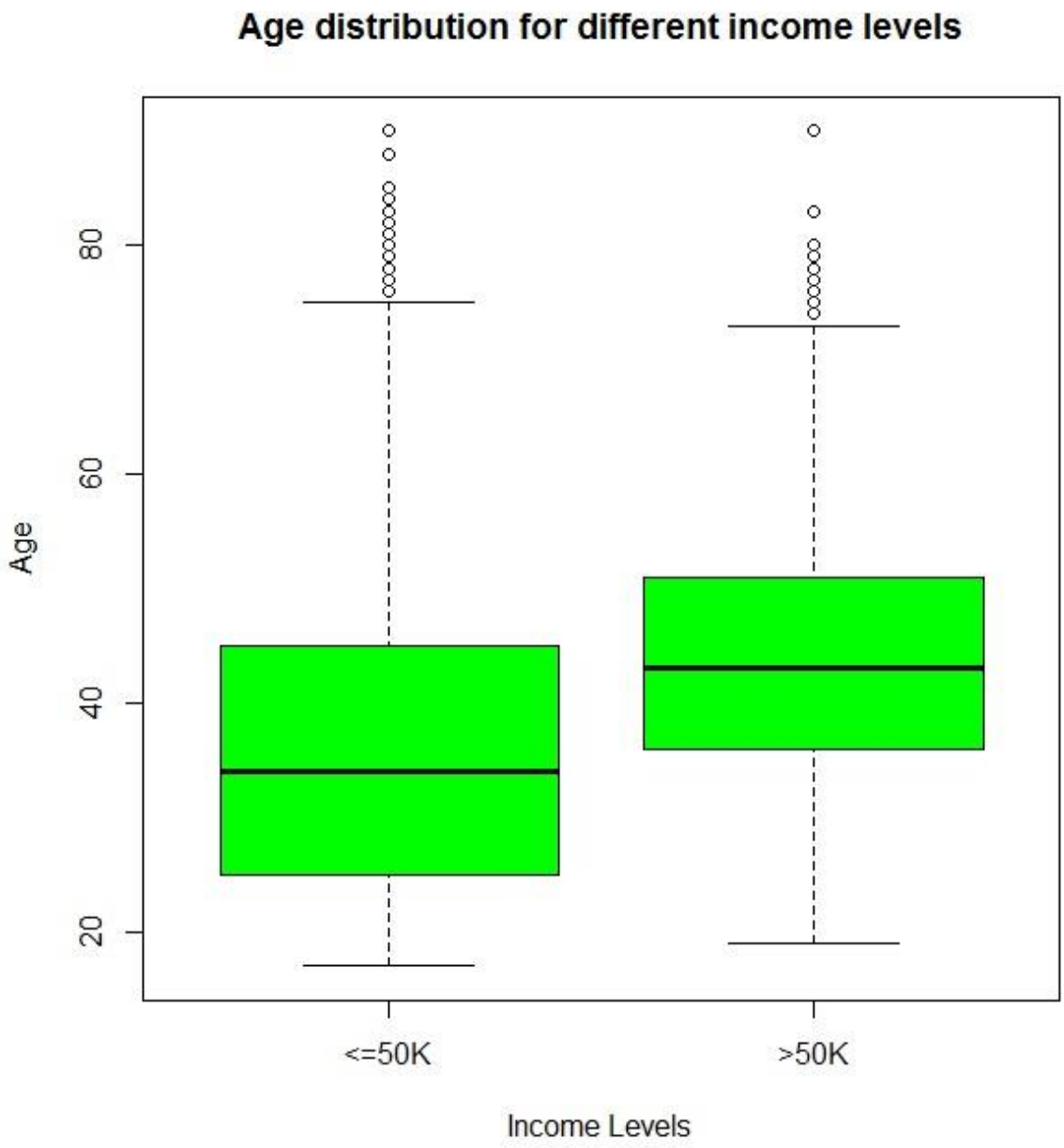
It is evident from the boxplots of age distributions over income levels for both male and female that more males above the age of forty earn more than $50000 than females.

2.5.3 Exploring Data by Age

The age distribution for income level less that $50,000 is skewed to the right with dropping proportion as age increases while age distribution for income level greater than $50,000 is symmetric and evenly distributed with highest proportion around age 40.


Age distribution for income level <=50K
1000
750
500
250
0
25
50
75
Age
Income_Level
<=50K


Age distribution for income level >50K
1000
750
500
250
0
25
50
75
Age
Income_Level
>50K

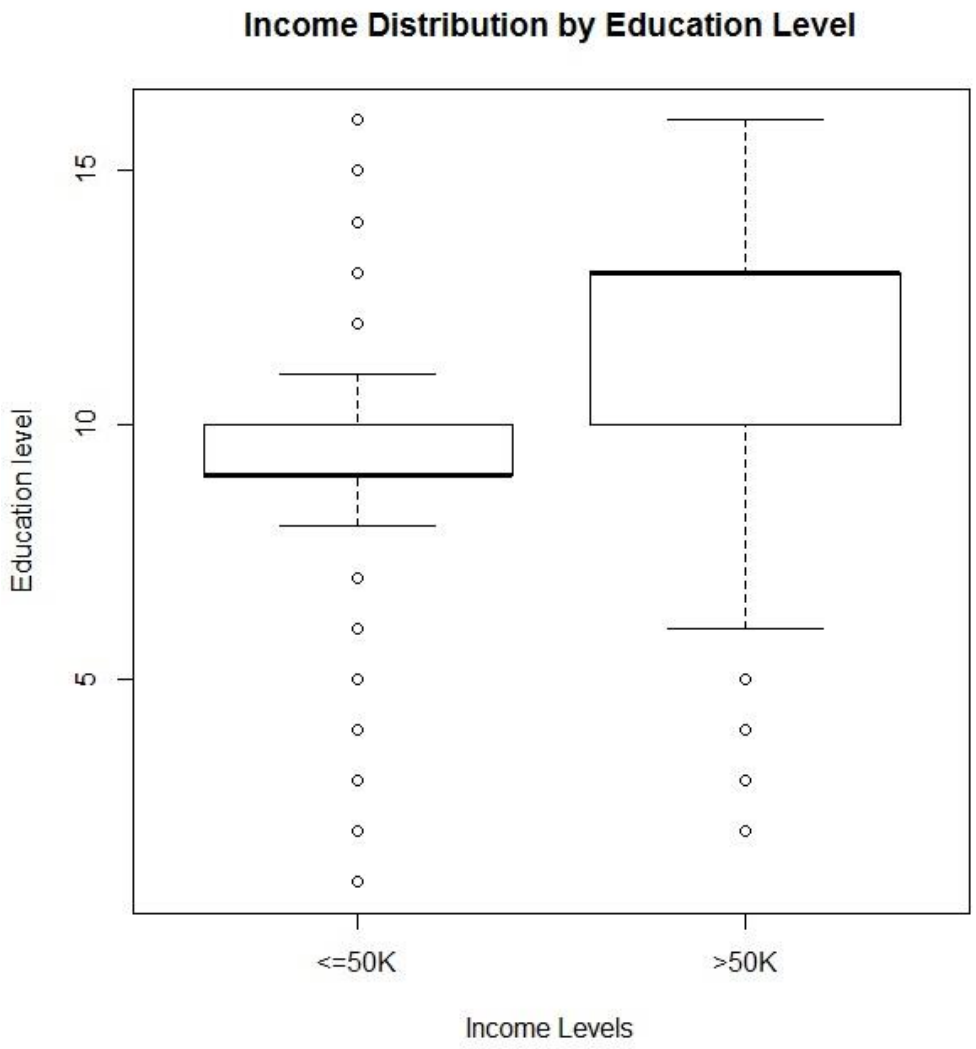


The age distribution for different income levels shows that the median age for those earning more than $50,000 is slightly above 40 while those earning less than or equal to $50,000 has median age lesser than 40.

2.5.4 Exploring Data by Education Level

Apparently, education level is one of the major indicators of income level. The median value for income levels greater than $50,000 is equivalent with the upper quartile of the distribution an indication that there is greater

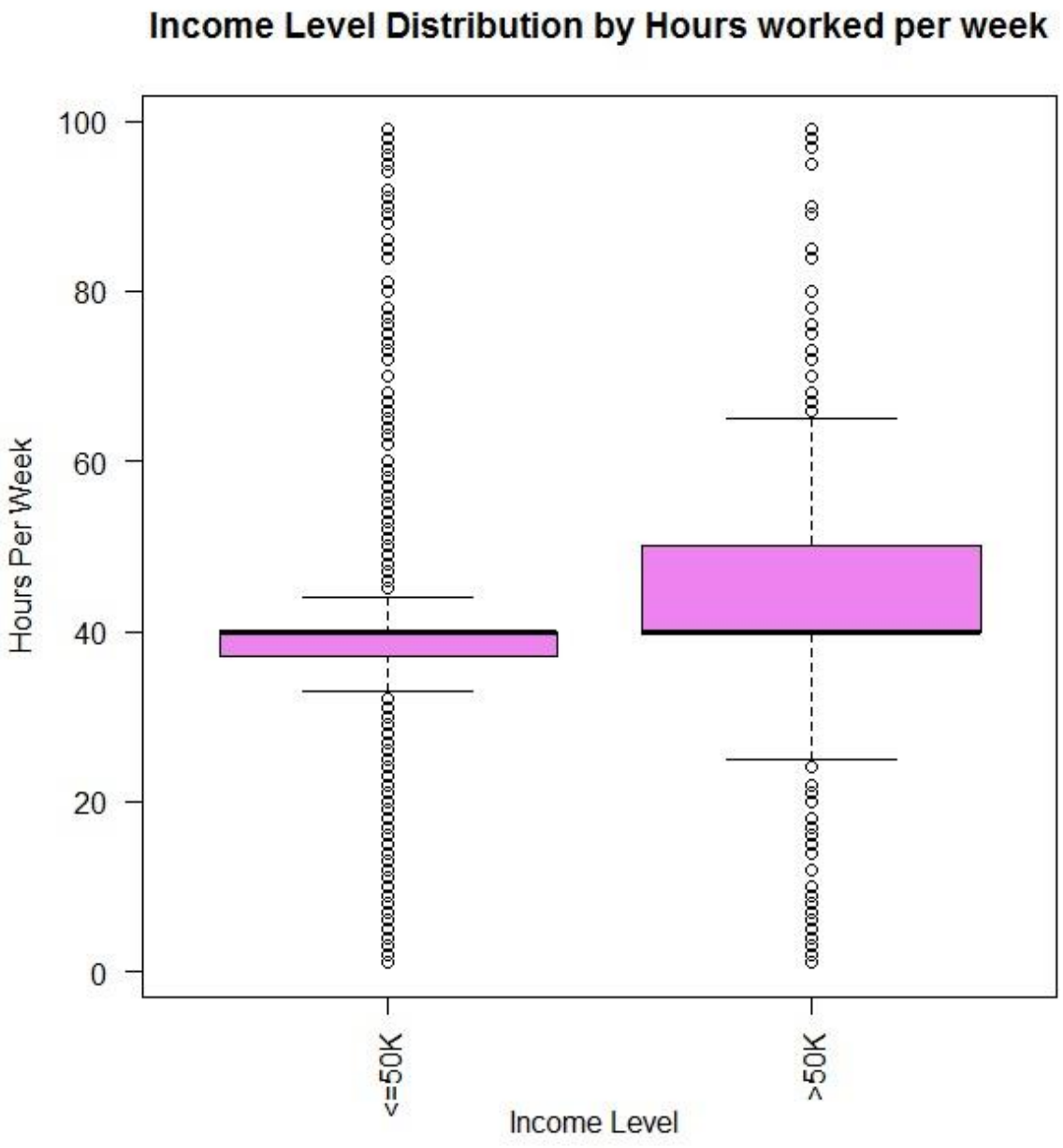
propensity for higher income as education level increases, while the median value for income levels lesser than $50,000 is equivalent to the lower quartile of the distribution, which also indicates that the lower the education level, the lesser the propensity for high income.



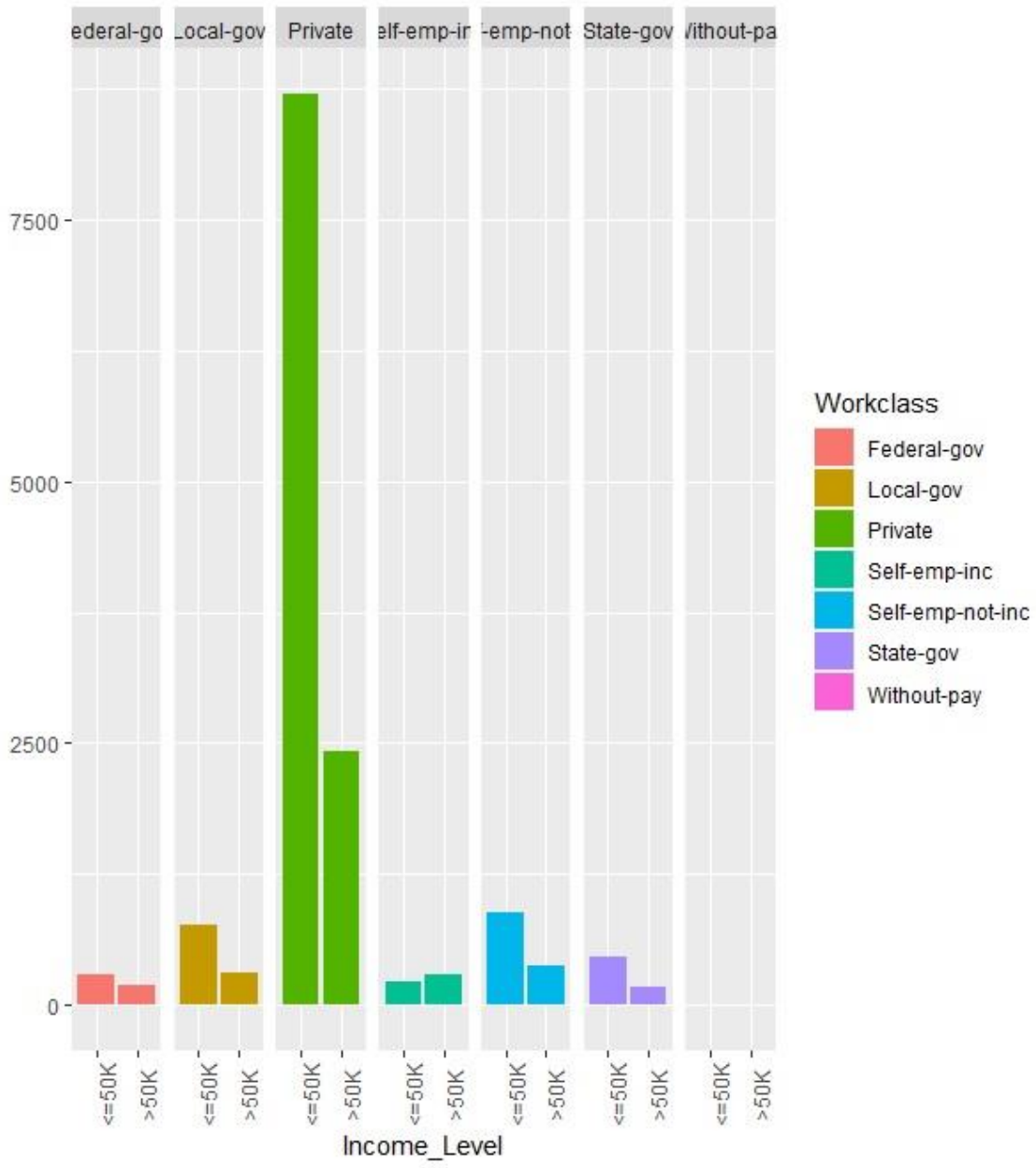
2.5.5 Exploring Data by Work Duration

The income level distribution by work duration for individuals earning less than or equal to $50,000 seem to concentrate around 40 hour per week also being its media value, meanwhile the work duration for individuals earning more than $50,000 concentrates between 25 to 65 hours per week. This is a clear indication that more work duration attracts more earning. On the

contrary, the median value for '>50K' distribution for hours per week coinciding with the lower quartile shows that more individuals earning over $50,000 work less. Moreover, both income levels '<=50K' and '>50K' have the same median value of 40 hours per week. Logically hours per week is not a good solitary indicator of income level.

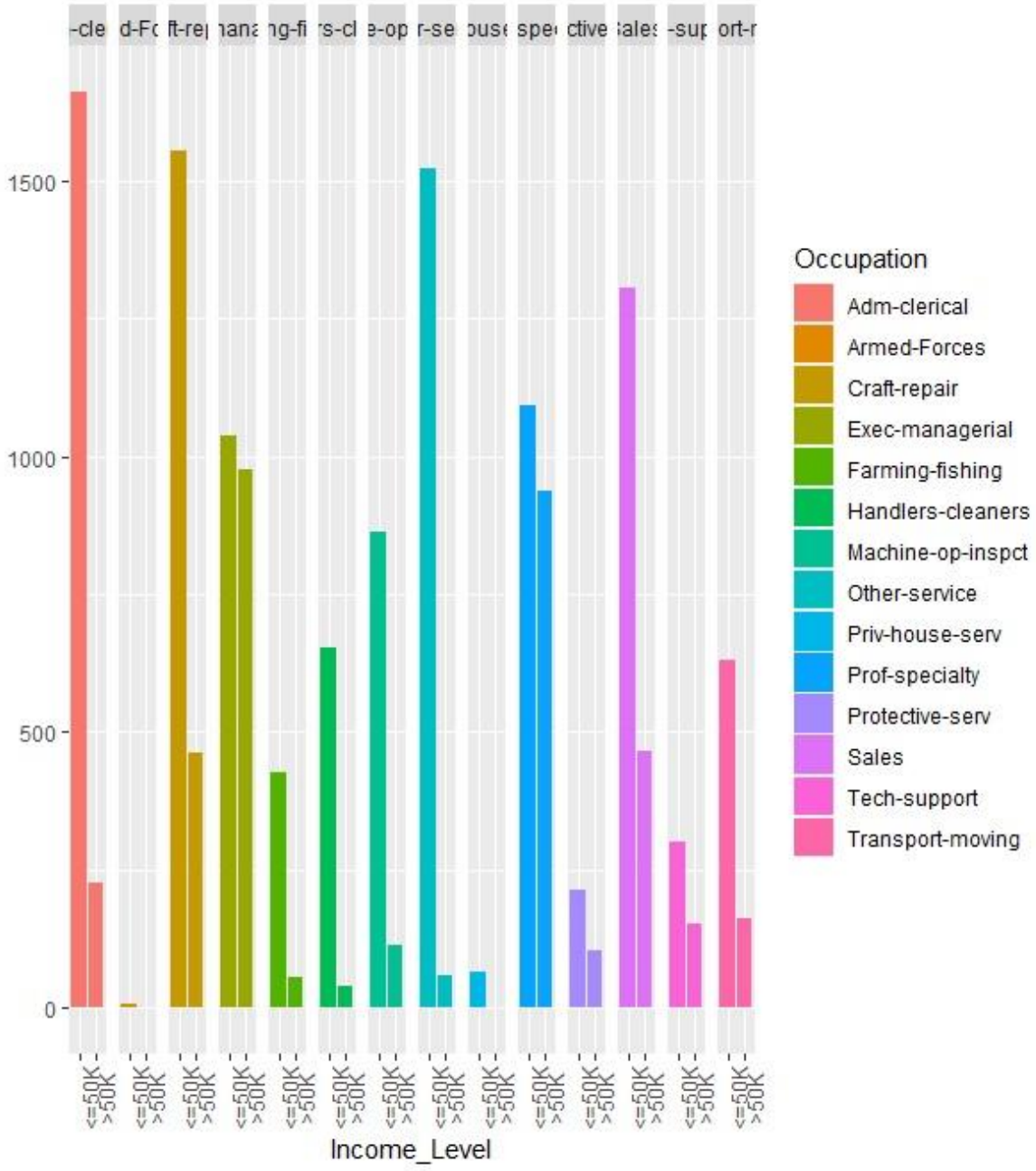


2.5.6 Exploring Data by Workclass



This reveals that there are absolutely more individuals in the private sector, implying a capitalist economy. Self-employed not incorporated is the second employer of labour. The Federal government seem to be the least employer of labour. Apparently, self-employed incorporated have more individuals that earn above $50,000 than those earning less than or equal to $50,000. What this goes to show is that creating a venture is more lucrative and a better pathway to economic liberation, social value and wellbeing.

2.5.7 Exploring Data by Occupation



Two occupations, Exec-managerial and Prof-specialty seem to similar pattern as there as roughly as much individual earning above $50,000 as those earning less than or equal to $50,000 despite the prevalence of "<=50K" income level. This shows that some occupation are more lucrative than others. Admin-clerical and other-service have much more individuals earning less than or equal to $50,000. Meanwhile, occupations such as Armed-forces and priv-house-serv have no individual earning above $50,000.

2.6 MODELLING

Building a prediction model requires some techniques, in this study the independent variables will be used except the Final_Weight variable to build and test a demographic model (Dem_model) that predicts an outcome, i.e. the income level of an individual to be greater than USD 50000 or less than USD 50000 using entire demographic data.

Secondly another model, the Human Capital model (HC_model) will be built and tested using most significant features of Human capital that predicts an outcome i.e. the income level of an individual to be greater than USD 50000 or less than USD 50000 using specific features from the demographic data.

The data frame contains all the data to be used. The formula has the form `[outcome ~ predictor_1 + predictor_2 + predictor_3]` and so on. Since most demographic data are categorical and apparently are weak predictors, the *Boosting algorithm will be required for classification modeling*. Cross Validation (CV) where the training data is partitioned a specific number of times is also applied and separate boosted models are built on each. The resulting models are cumulated to arrive at final model, this helps avoid overfitting the model to the training data.

2.6.1 Notation

In machine learning, data comes in the form of:

1. the *outcome* we want to predict and
2. the *features* that we will use to predict the outcome

We want to build an algorithm that takes feature values as input and returns a prediction for the outcome when we do not know the outcome. The machine learning approach is to *train* an algorithm using a dataset for

which we do know the outcome, and then apply this algorithm in the future to make a prediction when we do not know the outcome.

Here Y will be used to denote the outcome and X1,…,Xp to denote features. Note that features are sometimes referred to as predictors or covariates, which are considered to be synonyms.

Prediction problems can be divided into categorical and continuous outcomes. For categorical outcomes, Y can be any one of K classes. The number of classes can vary greatly across applications. For example, in the digit reader data, K=10 with the classes being the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. In speech recognition, the outcomes are all possible words or phrases we are trying to detect. Spam detection has two outcomes: spam or not spam. In this study, the categories are denoted with indexes k=1,…,K. However, for binary data we will use k=0,1 (which in this study is k= "<=50K", ">50K") for mathematical conveniences that we demonstrate later.

The general setup is as follows. There are series of features and an unknown outcome to predict:

Outcome	feature 1	feature 2	feature 3	feature 4	…	feature n
?	X1	X2	X3	X4	…	Xn

To *build a model* that provides a prediction for any set of observed values X1=x1, X2=x2,…Xn=xn, data with known outcomes are collected.

Outcome	feature 1	feature 2	feature 3	feature 4	feature n
y1	x1,1	x1,2	x1,3	x1,4	x1,n
y2	x2,1	x2,2x	x2,3	x2,4	x2,n
⋮	⋮	⋮	⋮	⋮	⋮
Ym	xm,1	xm,2	xm,3	xm,4	xm,n

When computing overall accuracy of the prediction algorithm, there might be high percentage of mistakes because the prevalent outcome may outweigh the other resulting in biased gain for the prevalent outcome. This can actually be a big problem in machine learning. If a training data is biased in some way, the algorithm is much more likely to develop bias as well. The fact that test set are used to validate results does not matter because it is also derived from the original biased dataset. This is one of the reasons why it is important to look at metrics other than overall accuracy when evaluating a machine learning algorithm.

2.6.2 Prevalence

Prevalence is the percentage proportion of the 'positive class' in the dataset that is being analyzed. For instance the outcome to be predicted in this study (income level) contains x numbers of outcome variable "<=50K" and y numbers of outcome variable ">50K" and the positive class is "<=50K", then prevalence can be calculated as mean of the positive class times 100. i.e. $x/(x+y) * 100$

However, because there is usually a high tendency for deception when estimating the overall accuracy due to prevalence, there are several metrics that can use to evaluate an algorithm in a way that prevalence does not cloud the assessment, and these can all be derived from the confusion matrix. A general improvement to using overall accuracy is to study *sensitivity* and *specificity* separately.

2.6.3 The Confusion Matrix

A confusion matrix is a table that describes the performance of a classifier/classification model. It contains information about the *actual and prediction classifications* done by the classifier and this information is used to evaluate the performance of the classifier.

Note that the confusion matrix is only used for classification tasks, and as such cannot be used in regression models or other non-classification models.

The prediction rule will be to predict an outcome (income levels) based on features in the dataset. This can be done by constructing the *confusion matrix*, which basically tabulates each combination of prediction and actual value. The confusion matrix can be visualized as below:

		PREDICTED	
		<=50k	>50K
ACTUAL	<=50k	1	0
	>50K	0	1

2.6.4 Sensitivity and Specificity

To define sensitivity and specificity, we need a binary outcome. When the outcomes are categorical, we can define these terms for a specific category. In the digits example, we can ask for the specificity in the case of correctly predicting 2 as opposed to some other digit. Once we specify a category of interest, then we can talk about positive outcomes, $Y=1$, and negative outcomes, $Y=0$.

In general, *sensitivity* is defined as the ability of an algorithm to predict a positive outcome when the actual outcome is positive: $Y=1$ when $Y=1$. Because an algorithm that calls everything positive ($\hat{Y}=1$ no matter what) has perfect sensitivity, this metric on its own is not enough to judge an algorithm. For this reason, we also examine *specificity*, which is generally defined as the ability of an algorithm to not predict a positive $\hat{Y}=0$ when the actual outcome is not a positive $Y=0$. We can summarize in the following way:

- High sensitivity: $Y = 1 \Longrightarrow \hat{Y} = 1$
- High specificity: $Y = 0 \Longrightarrow \hat{Y} = 0$

Although the above is often considered the definition of specificity, another way to think of specificity is by the proportion of positive calls that are actually positive:

- High specificity: $\hat{Y} = 1 \Longrightarrow Y = 1$

To provide precise definitions, we name the four entries of the confusion matrix:

	Actually Positive	**Actually Negative**
Predicted positive	True positives (TP)	False positives (FP)
Predicted negative	False negatives (FN)	True negatives (TN)

Sensitivity is typically quantified by TP/(TP+FN), the proportion of actual positives (the first column = TP+FN) that are called positives (TP). This quantity is referred to as the *True Positive Rate* (TPR) or *Recall*.

Specificity is defined as TN/(TN+FP) or the proportion of negatives (the second column = FP+TN that are called negatives (TN). This quantity is also called the true negative rate (TNR). There is another way of quantifying specificity which is TP/(TP+FP) or the proportion of outcomes called positives (the first row or TP+FP) that are actually positives (TP). This quantity is referred to as *Positive Predictive Value (PPV)* and also as *Precision*.

Note that, unlike TPR and TNR, precision depends on prevalence since higher prevalence implies you can get higher precision even when guessing. The multiple names can be confusing, so we include a table to help us remember the terms. The table includes a column that shows the definition if we think of the proportions as probabilities.

Measure of	**Name 1**	**Name 2**	**Definition**	**Probability representation**
Sensitivity	TPR	Recall	TP/(TP+FN)	$Pr(\hat{Y}=1\|Y=1)$
Specificity	TNR	1-FPR	TN/(TN+FP)	$Pr(\hat{Y}=0\|Y=0)$
Specificity	PPV	Precision	TP/(TP+FP)	$Pr(Y=1\|\ \hat{Y}=1)$

Here TPR is True Positive Rate, FPR is False Positive Rate, and PPV is Positive Predictive Value. The **caret** function [confusionMatrix] computes all these metrics once the "positive" category is defined. The function expects features as input, and the first level is considered the positive outcome or Y=1.

It is possible for the prediction algorithm to output high accuracy even when sensitivity or specificity are low, especially if the prevalence of the positive outcome is low. This is the main reason why it is important to examine sensitivity and specificity and not just accuracy, thus before applying an algorithm to general datasets, there is need to find out whether prevalence will be the same.

CHAPTER THREE

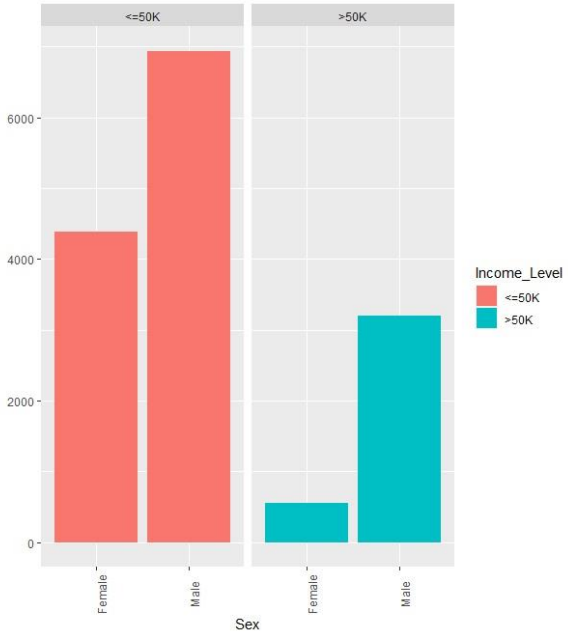
RESULT

The model result shows the product of the machine learning algorithm that is designed to predict individual's income level given that there are two classes of income levels: income greater than $50,000 (>50K) and income lesser or equal to $50,000 (<=50K). The model uses the confusion matrix to analyze significant data believed to be predictors of income. The ultimate aim of human capital development is to increase income, which arguably improves wellbeing. Two strong predicators, gender and education level were examined in the first phase.

3.1 Income level prediction by gender

```
   Acc   Sen   spe  prev
 <dbl> <dbl> <dbl> <dbl>
 0.503 0.888 0.316 0.672
```

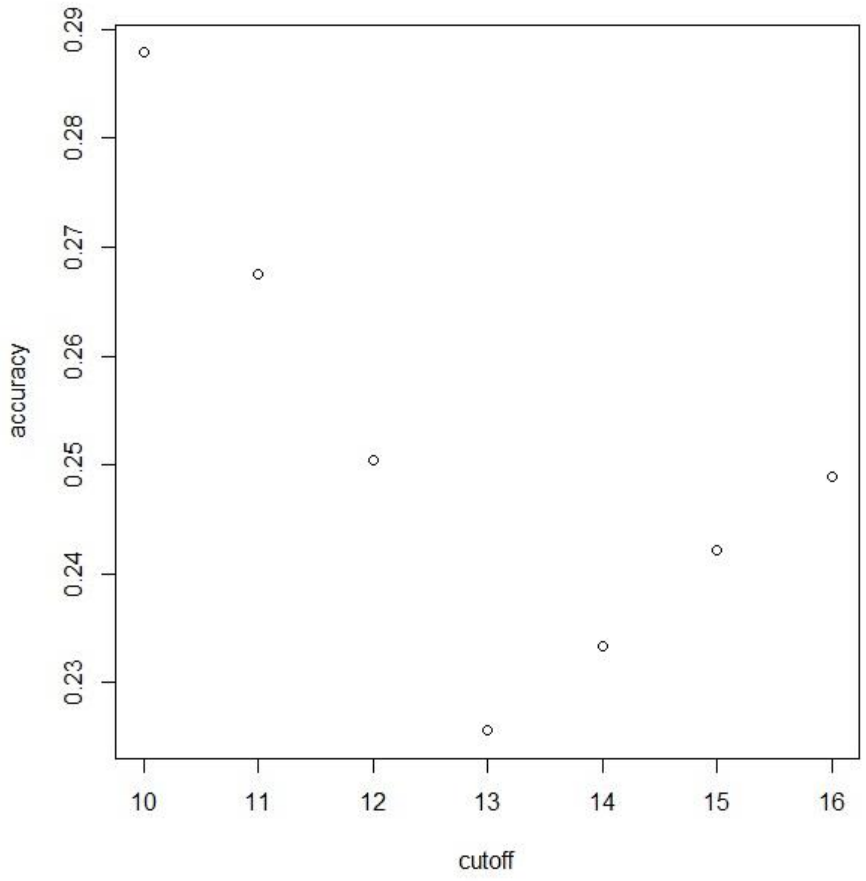
The result above indicates that the model predicting a male to earn ">50K" is 50.3% accurate, with 88.8% sensitivity and 31.6% specificity. Though the male gender has 67.2% prevalence. Conversely, the female gender would be 49.7% accurate, which indicates that income level is not gender based as there is slight difference despite the prevalence.



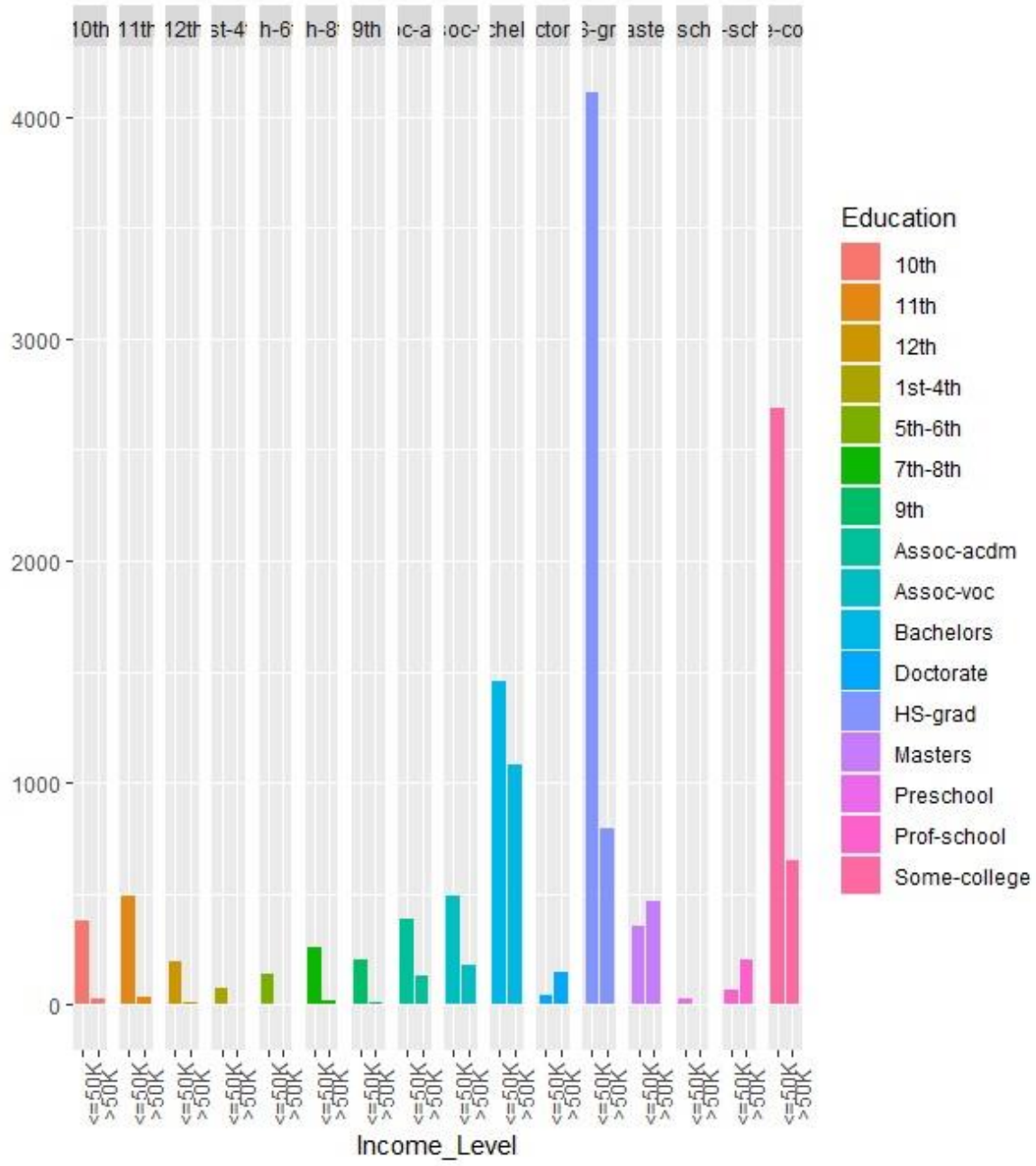
3.2 Income level prediction by education level

```
acc_Ed sen_Ed spe_Ed prev_Ed
 <dbl>  <dbl>  <dbl>   <dbl>
0.288  0.246  0.414   0.751
```

This model predicts income level for education number greater than the average. The prediction that individuals having more than the average education number earns "<=50K" is 28.8% accurate with 24.6% sensitivity and 41.4% specificity. Though the "<=50K" is 75.1% prevalent in the income level. Ultimately, it can be deduced that the prediction for individuals having more than average education number earning ">50K" is 71.2% accurate. Thus higher education will result in better income. This is the main crux of human capital development.



Accuracy was tested for values between the average education number to maximum (10 – 16) and the best cutoff produced this accuracy `0.2924209`.



As evident in the graph above, prof-school, master and doctorate has more individuals earning more than “>50K” than those earning “<=50K” while those with bachelor have a higher proportion of those earning “>50K” compared to lesser education ratings. This is a clear indication that education is a strong indicator for income level.

3.3 Model result using entire demographic data (Dem_Model)

The training set was subjected to the Confusion matrix model. In this case all the demographic details were included as predictors in the model except the *Final_Weight* data. The goal is to analyze all the predictors in the algorithm and predict income levels "<=50K" or ">50K" based on available data.

```
          Reference
Prediction <=50K  >50K
     <=50K 10725   602
     >50K   1405  2349

               Accuracy : 0.8669
                 95% CI : (0.8614, 0.8723)
    No Information Rate : 0.8043
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6167

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8842
            Specificity : 0.7960
         Pos Pred Value : 0.9469
         Neg Pred Value : 0.6257
             Prevalence : 0.8043
         Detection Rate : 0.7112
   Detection Prevalence : 0.7511
      Balanced Accuracy : 0.8401

       'Positive' Class : <=50K
```

This result shows that using these demographic details `(Age + Workclass + Education + Education_num + Marital_Status + Occupation + Relationship + Sex + Race + Capital_Gain + Capital_Loss + Hours_Per_week + Native_Country)` to predict income level that is "<=50K) is 86.7% accurate with 88.4% sensitivity, 79.6% specificity. Meanwhile the prevalence of the positive class (<=50K) is 80.4%

Demographic model (Dem_model) testing

The demographic data model was tested using the test_set and the result is as follows:

```
          Reference
Prediction <=50K  >50K
     <=50K 10659   668
     >50K   1468  2286

               Accuracy : 0.8584
                 95% CI : (0.8527, 0.8639)
    No Information Rate : 0.8041
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5922

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8789
            Specificity : 0.7739
         Pos Pred Value : 0.9410
         Neg Pred Value : 0.6090
             Prevalence : 0.8041
         Detection Rate : 0.7068
   Detection Prevalence : 0.7511
      Balanced Accuracy : 0.8264

       'Positive' Class : <=50K
```

The model testing for prediction using the entire demography produced accuracy of 85.8% with 87.9% sensitivity and 77.4% specificity. These parameters validates the result from the training data for the Dem_model with the accuracy falling by approximately 1%.

3.4 Human Capital Model (HC_Model)

[Result of Modeling with most significant Human Capital Indicators]

Human Capital is mostly indicated by the following data fields (`Age + Education + Education_num + Sex + Workclass + Occupation + Hours_Per_week`) according to the available data. Thus building a model with these significant fields instead of the entire demographic details produces the following result

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K  >50K
     <=50K 10538   789
     >50K   1910  1844

               Accuracy : 0.821
                 95% CI : (0.8148, 0.8271)
    No Information Rate : 0.8254
    P-Value [Acc > NIR] : 0.9228

                  Kappa : 0.4683

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8466
            Specificity : 0.7003
         Pos Pred Value : 0.9303
         Neg Pred Value : 0.4912
             Prevalence : 0.8254
         Detection Rate : 0.6988
   Detection Prevalence : 0.7511
      Balanced Accuracy : 0.7735

       'Positive' Class : <=50K
```

The model predicted income level with 82.1% accuracy with 84.7% sensitivity and 70% specificity. The positive class (<=50K) has 82.5% prevalence. These parameters tend to produce more accurate result having weaned out the weak predictors in the dataset.

Final Testing

[Human Capital Model (HC_Model) testing]

```
Confusion Matrix and Statistics

          Reference
Prediction <=50K  >50K
     <=50K 10502   825
     >50K   2009  1745

               Accuracy : 0.8121
                 95% CI : (0.8058, 0.8183)
    No Information Rate : 0.8296
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4382

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8394
            Specificity : 0.6790
         Pos Pred Value : 0.9272
         Neg Pred Value : 0.4648
             Prevalence : 0.8296
         Detection Rate : 0.6964
   Detection Prevalence : 0.7511
      Balanced Accuracy : 0.7592

       'Positive' Class : <=50K
```

The final model testing for predicting income level using the most significant human capital features or indices produces 81.2% accuracy with 83.9% sensitivity and 67.9% specificity. The positive class (<=50K) has 83%. These parameters also validates the result from the training data for the HC_Model with the accuracy falling by approximately 1% as well.

However, due to the nature of the dataset, most of the variables are categorical dataset. Apparently, non-parametric data are weak predictors. The demographic model (Dem_model) used 13 predictors in the algorithm and produced [0.8584] accuracy while the Human Capital model (HC_model) used 7 predictors and produced [0.8121] accuracy indicating that almost half of the predictor in the demographic model had little impact on the prediction process.

S/N	**Dem_model prediction features**	**HC_Model prediction features**
1	Age	Age
2	Workclass	Workclass
3	Education	Education
4	Education_num	Education_num
5	Marital_Status	
6	Occupation	Occupation
7	Relationship	
8	Sex	Sex
9	Race	
10	Capital_Gain	
11	Capital_Loss	
12	Hours_Per_week	Hours_Per_week
13	Native_Country	

Meanwhile, there seem to be little or no correlation between the continuous variables in the dataset.

	Age	Education_num	Hours_Per_week
Age	1.0000000	0.0491466	0.1035191
Education_num	0.0491466	1.0000000	0.1455869
Hours_Per_week	0.1035191	0.1455869	1.0000000

CHAPTER FOUR

4.1 SUMMARY AND CONCLUSION

Human capital cannot be absolutely quantified because it is more or less an intangible feature which gives an entity the edge over its competitors. Human capital development is the process of improving those qualities that stands out an entity from the pack. Human capital development could be individual base (a person) or collectively (as a nation). A country may be ranked higher/lower with a counterpart nation in terms of human capital, while individuals could also be assessed in terms of human capital. Improving human capital increases the prospect of productivity, which invariably improves income and in the long run impact socio-economic progress which arguably is a panacea for wellbeing.

It is a common believe that human capital is primarily assessed based on gender and education level even though other features are equally significant. Thus this study examined the gender and education using baseline prediction algorithms. It was discovered that income level is not gender based because the baseline prediction accuracy for “Male” earning a “>50K” is 50.3%, while the baseline prediction accuracy for “Female” earning a “>50K” is 49.7 even though the male is nearly 75% more prevalent than female.

Meanwhile, the baseline prediction for education level shows that individuals having more than average education number and earning “<=50K” is 28.8% while the baseline prediction that individuals having more than the average education number earning “>50K” is 71.2% accurate, which means income level is directly affected by education level.

Conclusively, a prediction model was built with the entire demographic (census) data [dem_model] using boosting algorithm which performed better than the model of the Human capital features (most significant features of human capital) [HC_model]. That is, when there are more useful predictors involved in model building irrespective of their impact, the propensity to predict correctly increases.

4.2 LIMITATIONS OF THE STUDY

The major limitation of this study is the scope of the data. The data is typically a census conducted in the United States of America, and out of the 32,561 rows, 29,170 represents individuals in the United States of America, which means approximately 90% of the data set represent a particular native country, while the remaining 10% represent the others. Ultimately, the findings might be based on what is obtainable in the United States of America which may differ when a more evenly distributed data is applied.

Secondly, since the survey was conducted in 1994, by 26 years later the data may have been stale as income levels and other features of the data may have drastically changed in connection to human capital development.

Lastly, most features of the dataset are categorical data. Categorical or nonparametric data are weak predictors.

4.3 FUTURE WORK

In the future, it is recommended that recent and balanced dataset are applied to studies of this magnitude. Also where necessary and practicable nonparametric data should be given numeric values in the order of precedence as this will improve data or model performance because during data analysis, it is very important to understand how the built model has performed with respect to existing dataset. This will help to understand the significance of the model when applied to data at real time.

REFERENCES

Carmignani F. and Chowdhury A. (2011). Four Scenarios of Development and the Role of Economic Policy
Available at: https://doi.org/10.1080/00220388.2010.506920

Kenton W, And Sonnenshein M. (2020). Human Capital. Investopedia.
Available at: https://www.investopedia.com/terms/h/humancapital.asp

Rafael A. Irizarry (2019). Introduction to Data Science: Data Analysis and Prediction Algorithms with R

Witten I. H., Frank E., Hall M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Third Edition. Morgan Kaufmann Series, USA. Elsevier, ISBN 978-0-12-374856-0