Adult_Census_Income

Arantxa Sanchis

21/06/2020

1. Executive Summary

Increasing amount of data in the rapidly expanding technological world of today makes the analysis of it much more exciting. The insights gathered from user data is now a major tool for the decision-makers.

Machine learning algorithms enable computers to learn from data, and even improve themselves, without being explicitly programmed. Machines are getting more and more intelligent and AI is expanding to more businesses and industries. With large-scale data available, scientists have started to build intelligent systems that are able to analyze and learn from large amounts of data.

This project study is in the field of income inequality. As nations progress it is experienced that the income differential between the rich and the poor or for that matter between various classes of population, widens. In order to address the discrimination it is imperative that governments collect data and analyze the same. Accordingly, the basic aim of this study is to use the dataset of "Adult Census Income" and apply machine learning and data mining techniques to suggest a solution to the income inequality problem.

1.1 Introduction

The glaring inequality of wealth and income is a huge concern especially in the United States. The chances of reducing poverty can be spurred by tackling the surging level of economic inequality in the world. The concept of equality ensures sustainable development and economic stability of a nation. Different countries have been trying their best to alleviate this problem by extensive studies leading to near optimal solutions.

Classification will be done to predict whether an individual's yearly income in US falls in the income category of either greater than $50K or less than equal to $50K based on a certain set of attributes.

1.2 Goal of the project

The objective of this project is to use various machine learning algorithms to build efficient models assisted by teachings in the Harvard Course – Data Science & Machine Learning to predict whether an individual's income is greater than $50k or less than or equal to $50k whilst considering the effect of other variables appearing in the Adult Census Income dataset.

1.3 Approach Used

Our machine learning algorithm will be evaluated based on the accuracy of our predictions made to the "validation" set. We will explore several combinations of features and predictors to train different models in order to improve the overall accuracy.

Our aim is to find the model that gives us the highest accuracy.

This project analyses the dataset using four different machine learning algorithms.

1) K Nearest Neighbours (KNN)

K-nearest neighbors is a non-parametric method used for classification and regression. The basic logic behind KNN is to explore your neighborhood, assume the test data point to be similar to them and derive the output. In KNN, we look for k neighbors and come up with the prediction.

In case of KNN classification, a majority voting is applied over the k nearest data points whereas, in KNN regression, mean of k nearest data points is calculated as the output. As a rule of thumb, we select odd numbers as k.

KNN is a lazy learning model where the computations happens only at run time. It is one of the most easy machine learning techniques used. It is a lazy learning model, with local approximation.

It works by calculating the distance between observations based on the attributes. New data points, or observations, are predicted by looking at the k-nearest points and averaging them. Therefore, if the majority of K-neighbors belong to a certain class, the new observation also belongs to that same class.

2) Classification and Regression Trees (CART)

A Classification And Regression Tree (CART) is a predictive model, which explains how an outcome variable's values can be predicted based on other values. A CART output is a decision tree where each fork is a split in a predictor variable and each end node contains a prediction for the outcome variable.

3) Gradient Boosting Machines (GBM)

GBMs build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined, these many weak successive trees produce a powerful "committee".

The main idea of boosting is to add new models to the ensemble sequentially. At each particular iteration, a new weak, base-learner model is trained with respect to the error of the whole ensemble learnt so far.

4) Random Forests (RF)

The random forest is a classification algorithm consisting of many decisions trees in order to improve the predictions. It uses bagging/bootstrap and feature randomness when building each individual tree to try to create an uncorrelated ensemble of decision trees namely a "forest" of trees whose prediction by committee is more accurate than that of any individual tree. In general, it builds multiple decision trees and merges them together to get a more accurate and stable prediction. The general idea of the bagging method is that a combination of learning models increases the overall result.

They sample "N" observations with replacement from the training set to create a bootstrap training set. Another way that Random Forests introduce randomness is that each tree is built from its own randomly selected subset of features. This helps reduce the correlation between the trees. Finally, the Random Forest algorithm creates an ensemble by averaging the predictions of all the trees to form a final prediction.

1.4 Key Steps

Steps to build the Adult Census Income prediction model are broadly summarized as below.

1) Obtain the original "Adult Census Income" dataset and explore and analyze the data to study the features of the dataset. If required remove columns which are irrelevant and not required for the study.

2) Split the "Adult Census Income" dataset into two subsets-

a) "adult_census_income_training": a training subset to train the algorithm
b) "adult_census_income___validation" : a test subset to assess the accuracy of the best fitted model

This is done in order to accurately predict the income of the population.

3) Further divide the "adult_census_income_training" dataset into subsets as below-

a) a training set "train_set"
b) a test set "test_set"

4) Train a number of models on the "train_set" dataset and test the models first on the "test_set" dataset. Models can be varied based on features/effects.

NOTE: The validation data will NOT be used for training the algorithm and will ONLY be used for evaluating the accuracy of the final algorithm.

5) Predict the income and compute the accuracy for each model to find the best model.

6) We can use the best model to generate the final prediction on the "adult_census_income___validation" dataset and ascertain the final accuracy.

1.5 Dataset

The "Adult Census Income" dataset is an extract from the 1994 Census and contains 32,561 rows, each representing an individual person. It contains 15 columns that represent socio-economic factors, such as age, education, marital status, race etc., of census correspondents.

The "income" field in the database gives the annual income of the correspondent. This column will assist in determining whether a group of correspondents earn an annual salary of either greater than or less than or equal to $50K.

We are using the publicly available "Adult Census Income" dataset from the UCI Machine Learning Repository on "kaggle" as below:

https://www.kaggle.com/uciml/adult-census-income

We have uploaded this dataset to "Google Drive" and made it "publicly available to all". We will download it from "Google Drive" into a local folder on our system using the code below. We can load the dataset into R as below:

```
#install packages
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(dplyr)) install.packages("dplyr", repos = "http://cran.us.r-project.org")
if(!require(rpart)) install.packages("rpart", repos = "http://cran.us.r-project.org")
if(!require(randomForest)) install.packages("randomForest", repos = "http://cran.us.r-project.org")
if(!require(googledrive)) install.packages("googledrive", repos = "http://cran.us.r-project.org")
if(!require(httpuv)) install.packages("httpuv", repos = "http://cran.us.r-project.org")

#Deauthorize i.e do not request for any login credentials
drive_deauth()
drive_user()

#Download dataset from Google drive
```

```
downloaded_file_aci <- drive_download(as_id("1ic_BT3GVKn_pQy0f81ITQKPiEAV7QMVW"), overwrite = TRUE)
google_file_aci <- downloaded_file_aci$local_path

#Read dataset into RStudio
adult_census_income_dataset<-read.csv(google_file_aci)
```

We can see the number of rows and columns in the dataset as below:

```
dim(adult_census_income_dataset)
```

```
## [1] 32561    15
```

2. Analysis

2.1 Data Cleaning

We will first inspect the dataset for missing values. We notice these have been represented by "?" in the observations.

```
missing_check<- filter(adult_census_income_dataset,
                       workclass == "?"|occupation == "?"|native.country == "?")

nrow(missing_check)
```

```
## [1] 2399
```

```
head(missing_check)
```

```
##   age workclass fnlwgt    education education.num     marital.status
## 1  90         ?  77053      HS-grad             9            Widowed
## 2  66         ? 186061 Some-college            10            Widowed
## 3  41   Private  70037 Some-college            10      Never-married
## 4  51         ? 172175    Doctorate            16      Never-married
## 5  22   Private 119592   Assoc-acdm            12      Never-married
## 6  61         ? 135285      HS-grad             9 Married-civ-spouse
##          occupation  relationship  race    sex capital.gain capital.loss
## 1                 ? Not-in-family White Female            0         4356
## 2                 ?     Unmarried Black Female            0         4356
## 3      Craft-repair     Unmarried White   Male            0         3004
## 4                 ? Not-in-family White   Male            0         2824
## 5 Handlers-cleaners Not-in-family Black   Male            0         2824
## 6                 ?       Husband White   Male            0         2603
##   hours.per.week native.country income
## 1             40  United-States  <=50K
## 2             40  United-States  <=50K
## 3             60              ?   >50K
## 4             40  United-States   >50K
## 5             40              ?   >50K
## 6             32  United-States  <=50K
```

A total of 2,399 rows have missing values in the dataset. We will exclude them as below:

```
adult_census_income_dataset <- filter(adult_census_income_dataset,
                                !workclass == "?",!occupation == "?",!native.country == "?")
adult_census_income_dataset <- droplevels(adult_census_income_dataset)
```

2.2 Data Analysis

We can get a glance of the dimensions and the first six rows of the dataset as below:

```
dim(adult_census_income_dataset)
```

```
## [1] 30162    15
```

```
head(adult_census_income_dataset)
```

```
##   age workclass fnlwgt    education education.num marital.status
## 1  82   Private 132870      HS-grad             9        Widowed
## 2  54   Private 140359      7th-8th             4       Divorced
## 3  41   Private 264663 Some-college            10      Separated
## 4  34   Private 216864      HS-grad             9       Divorced
## 5  38   Private 150601         10th             6      Separated
## 6  74 State-gov  88638    Doctorate            16  Never-married
##          occupation   relationship  race    sex capital.gain capital.loss
## 1   Exec-managerial  Not-in-family White Female            0         4356
## 2 Machine-op-inspct      Unmarried White Female            0         3900
## 3    Prof-specialty      Own-child White Female            0         3900
## 4     Other-service      Unmarried White Female            0         3770
## 5      Adm-clerical      Unmarried White   Male            0         3770
## 6    Prof-specialty Other-relative White Female            0         3683
##   hours.per.week native.country income
## 1             18  United-States  <=50K
## 2             40  United-States  <=50K
## 3             40  United-States  <=50K
## 4             45  United-States  <=50K
## 5             40  United-States  <=50K
## 6             20  United-States   >50K
```

The data now contains 30,162 rows and 15 columns. We would need to carry out a check that the dataset is complete in all aspects using the "summary" function as below.

```
summary(adult_census_income_dataset)
```

```
##       age         workclass            fnlwgt         education
##  Min.   :17.00   Length:30162       Min.   :  13769   Length:30162
##  1st Qu.:28.00   Class :character   1st Qu.: 117627   Class :character
##  Median :37.00   Mode  :character   Median : 178425   Mode  :character
##  Mean   :38.44                      Mean   : 189794
##  3rd Qu.:47.00                      3rd Qu.: 237629
##  Max.   :90.00                      Max.   :1484705
##  education.num   marital.status      occupation        relationship
```

```
##  Min.   : 1.00   Length:30162       Length:30162       Length:30162
##  1st Qu.: 9.00   Class :character   Class :character   Class :character
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :10.12
##  3rd Qu.:13.00
##  Max.   :16.00
##      race               sex             capital.gain    capital.loss
##  Length:30162       Length:30162       Min.   :    0   Min.   :   0.00
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.00
##  Mode  :character   Mode  :character   Median :    0   Median :   0.00
##                                        Mean   : 1092   Mean   :  88.37
##                                        3rd Qu.:    0   3rd Qu.:   0.00
##                                        Max.   :99999   Max.   :4356.00
##  hours.per.week  native.country        income
##  Min.   : 1.00   Length:30162       Length:30162
##  1st Qu.:40.00   Class :character   Class :character
##  Median :40.00   Mode  :character   Mode  :character
##  Mean   :40.93
##  3rd Qu.:45.00
##  Max.   :99.00
```

We will also use the "str" function to view the class of the objects as below.

```
str(adult_census_income_dataset)
```

```
## 'data.frame':    30162 obs. of  15 variables:
##  $ age           : int  82 54 41 34 38 74 68 45 38 52 ...
##  $ workclass     : chr  "Private" "Private" "Private" "Private" ...
##  $ fnlwgt        : int  132870 140359 264663 216864 150601 88638 422013 172274 164526 129177 ...
##  $ education     : chr  "HS-grad" "7th-8th" "Some-college" "HS-grad" ...
##  $ education.num : int  9 4 10 9 6 16 9 16 15 13 ...
##  $ marital.status: chr  "Widowed" "Divorced" "Separated" "Divorced" ...
##  $ occupation    : chr  "Exec-managerial" "Machine-op-inspct" "Prof-specialty" "Other-service" ...
##  $ relationship  : chr  "Not-in-family" "Unmarried" "Own-child" "Unmarried" ...
##  $ race          : chr  "White" "White" "White" "White" ...
##  $ sex           : chr  "Female" "Female" "Female" "Female" ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 3900 3900 3770 3770 3683 3683 3004 2824 2824 ...
##  $ hours.per.week: int  18 40 40 45 40 20 40 35 45 20 ...
##  $ native.country: chr  "United-States" "United-States" "United-States" "United-States" ...
##  $ income        : chr  "<=50K" "<=50K" "<=50K" "<=50K" ...
```

As observed above, our dataset contains a total of 15 variables of which 9 are categorical and 6 are continuous.

2.2.1 Studying the variables

1) Age The age structure of a population affects a nation's key socioeconomic issues. Countries with young populations (high percentage under age 15) need to invest more in schools, while countries with older populations (high percentage ages 65 and over) need to invest more in the health sector. The age structure can also be used to help predict potential political issues. For example, the rapid growth of a young adult population unable to find employment can lead to unrest.

In order to have a more rational view of the population, we can segregate the age into 5 main groups -

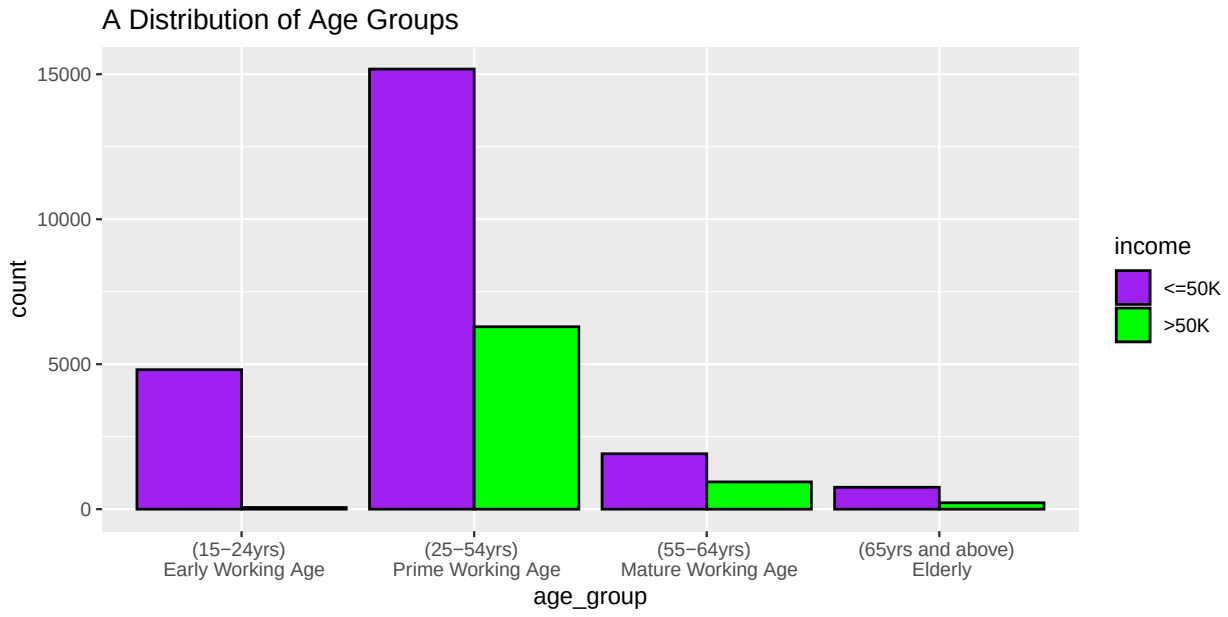
a) (0-14yrs) Children
b) (15-24yrs) Early Working Age
c) (25-54yrs) Prime Working Age
d) (55-64yrs) Mature Working Age
e) (65yrs and above) Elderly

We can group the age as below:

We will study the age group distribution of the census correspondents in the "Adult Census Income" dataset. We observe that majority of the population (71.2%) belongs to the "Prime Working Age" group. The "Early Working Age" group holds 16.1% of the population followed by the "Mature Working Age" group having 9.45% of the population.

As expected, the proportion of pensioners and retired people falling into the "Elderly" group are the lowest at 3.23%. No individuals below the age of 15 are working. It is observed that a large amount of variability exists in the age attribute. This appears to be an appropriate predictor for the income.

A study of the income differential reveals that, a very small fraction of the "Early Working Age" group population earn over $50K, as it requires adequate time and experience to grow professionally. It is observed that the proportion of individuals earning over $50K is less than half in comparison to those earning less than or equal to $50K in the remaining three groups - "Mature Working Age", "Mature Working Age" and "Elderly".



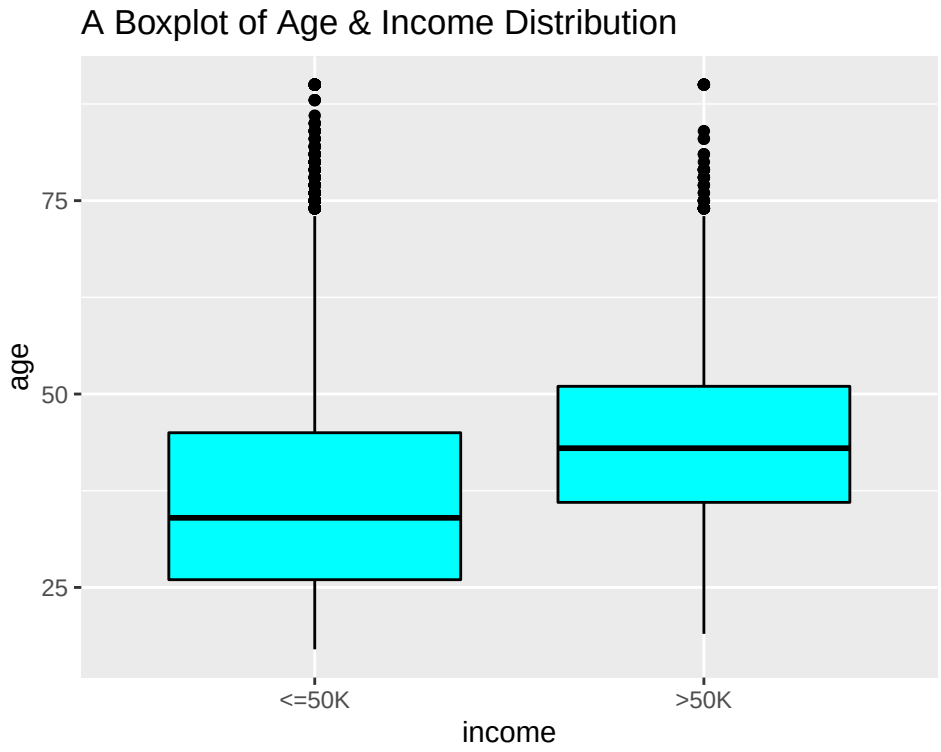
The percentage breakdown is as below:

```
## # A tibble: 4 x 3
##   age_group                              n percent_age
##   <chr>                              <int>       <dbl>
## 1 "(25-54yrs) \n Prime Working Age"  21469       71.2
## 2 "(15-24yrs) \n Early Working Age"   4869       16.1
## 3 "(55-64yrs) \n Mature Working Age"  2849        9.45
## 4 "(65yrs and above) \n Elderly"       975        3.23
```

We can see from the boxplot that the median age for individuals with greater than $50k income is considerably higher than the median age for individuals with less than or equal to $50k income. Also, the median tends

towards the center of the interval for those with over 50K income while it is nearer to the lower limit of the interval for those with less than or equal to 50K income.

```
adult_census_income_dataset %>%
ggplot(aes(income,age)) + geom_boxplot(color="black",fill="cyan") +
ggtitle("A Boxplot of Age & Income Distribution")
```



2) Working Class We observe that majority of the population (73.9%) belongs to the "Private" working class. The proportion of individuals having an income of over $50K in the "Private" working class is significantly higher when compared with other classes of over $50K income.

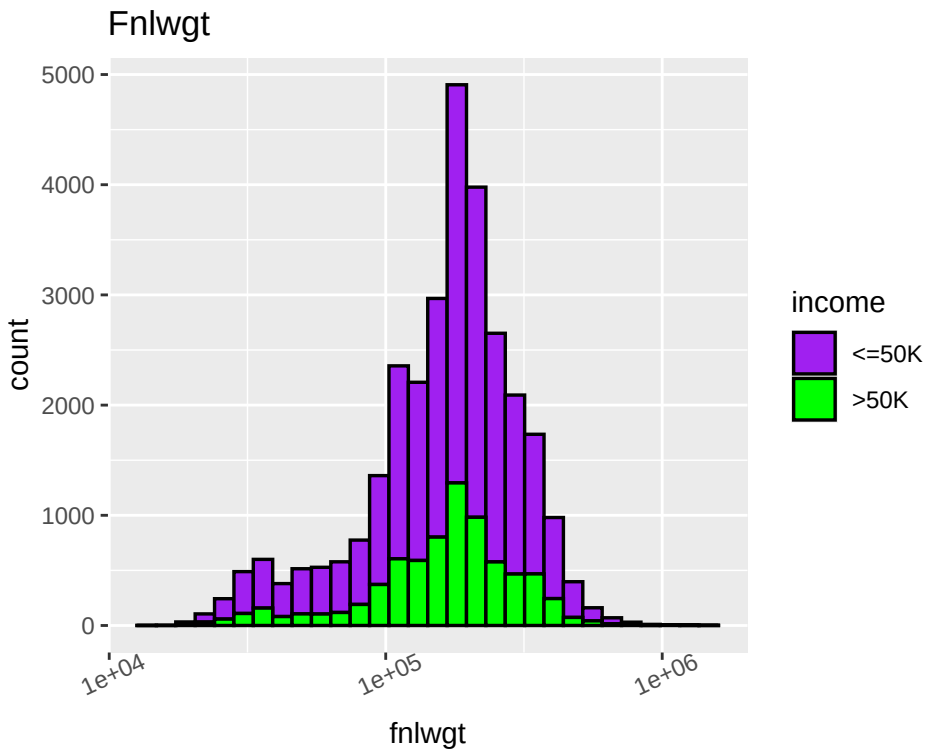
In all working classes we observe that the percentage of individuals earning less than or equal to $50K is substantially higher than those earning over $50K except for "Self-emp-inc" group where higher proportion of people belong to the over $50K income bracket. "Self-emp-inc" refers to people who work for themselves in corporate entities.

It is essential for governments to expand employment opportunities and promote self-sufficiency. Self-employment is a viable option and the general notion for many is a way to control their own futures and make work more fulfilling. Also because of the difficulty they have in finding employment many turn to self-employment to become self-sufficient. Even people with disabilities are capable of owning and operating a variety of businesses.

The percentage breakdown is as below:

```
## # A tibble: 7 x 3
##   workclass            n percent_workclass
##   <chr>            <int>             <dbl>
## 1 Private          22286             73.9
## 2 Self-emp-not-inc  2499              8.29
## 3 Local-gov         2067              6.85
## 4 State-gov         1279              4.24
## 5 Self-emp-inc      1074              3.56
## 6 Federal-gov        943              3.13
## 7 Without-pay         14              0.0464
```

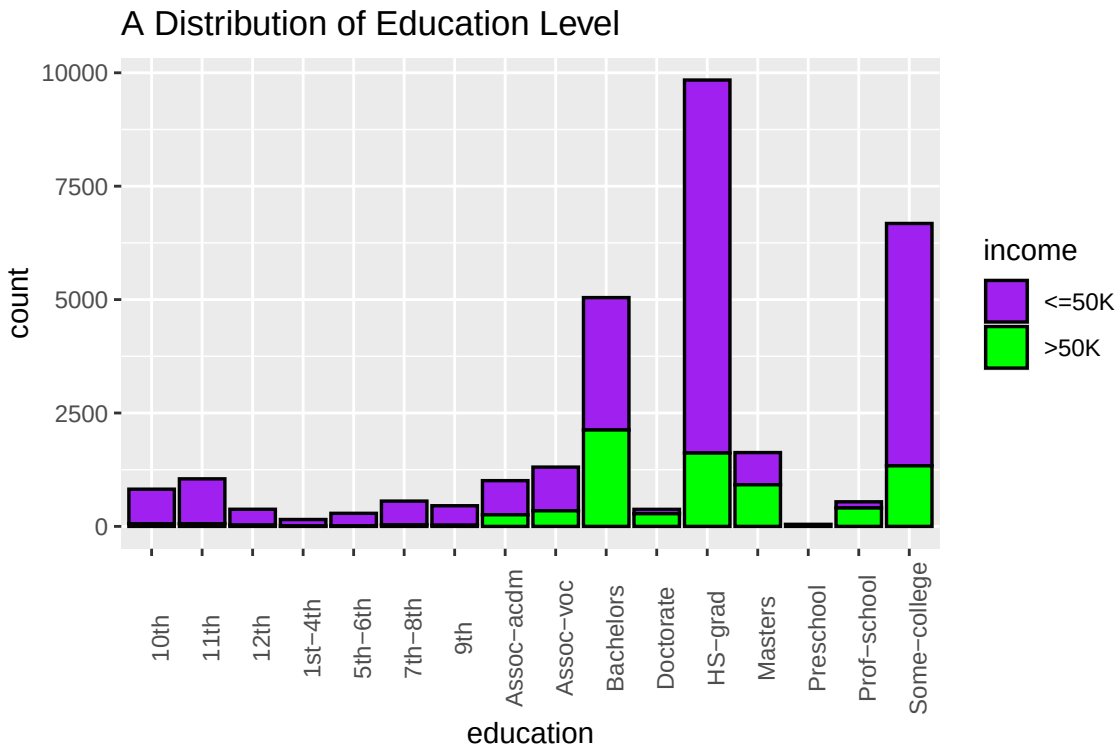
3) fnlwgt The "fnlwgt" attribute implies final weight, which is the number of units in the target population that the responding unit represents. In simple terms it is the number of people the census believes the entry (each data record) represents.



4) Education We observe a major proportion of the population (32.6%) are high school graduates and a vast majority of them earn less than or equal to $50K. As anticipated we see that a college degree significantly improves one's employment prospects and earning potential. Also despite possessing a Bachelor's degree it is seen that the proportion of individuals with over $50K income and those with less than or equal to $50K income is almost the same.

More individuals earn over $50K with the attainment of Prof-School, Masters or Doctorate degrees since professions with higher education and training requirements tend to pay workers higher wages. However, only a small proportion of individuals of the overall population enroll for these degrees.

A country's economy becomes more productive as the proportion of educated workers increase as they can more efficiently carry out tasks that require literacy and critical thinking. However, obtaining a higher level of education also carries a cost. The governments should provide basic literacy programs and also increase funding of higher studies to see faster economic growth and improved economic performance.



The percentage breakdown is as below:

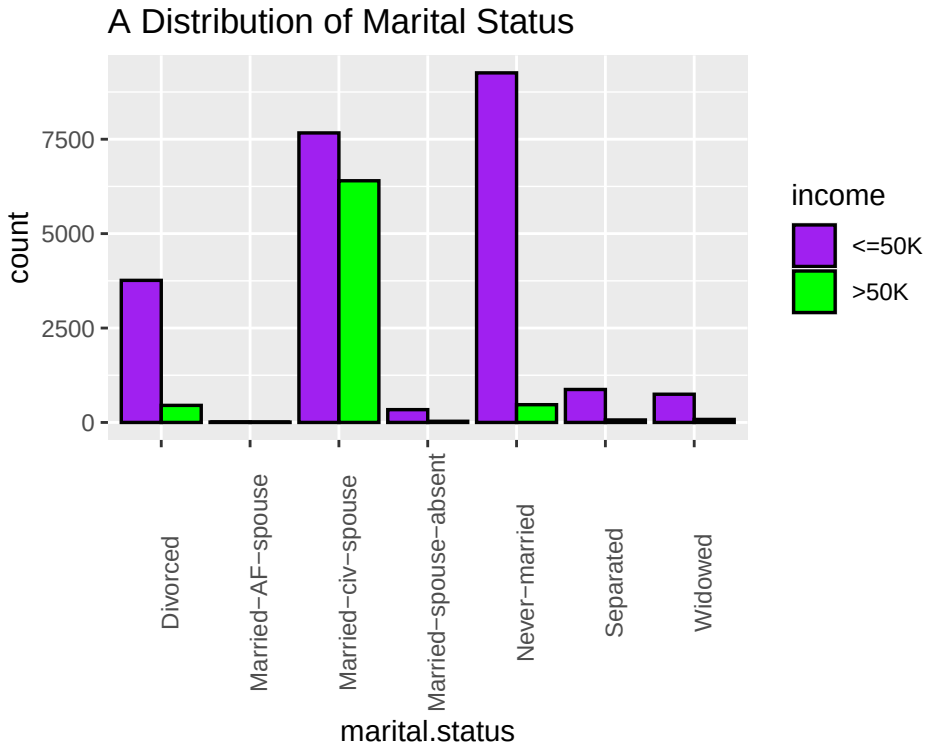
```
## # A tibble: 16 x 3
##    education        n percent_education
##    <chr>        <int>             <dbl>
##  1 HS-grad       9840              32.6
##  2 Some-college  6678              22.1
##  3 Bachelors     5044              16.7
##  4 Masters       1627               5.39
##  5 Assoc-voc     1307               4.33
##  6 11th          1048               3.47
##  7 Assoc-acdm    1008               3.34
##  8 10th           820               2.72
##  9 7th-8th        557               1.85
## 10 Prof-school    542               1.80
## 11 9th            455               1.51
## 12 12th           377               1.25
## 13 Doctorate      375               1.24
## 14 5th-6th        288               0.955
## 15 1st-4th        151               0.501
## 16 Preschool       45               0.149
```

5) Education num The education.num is simply a numerical representation of the education attribute. It ranges from 1 to 16 with 1 being the lowest (Preschool) and 16 being the highest (Doctorate).

6) Marital Status It is a well-known that two-parent families fare better financially than one-parent families. On observation, the dataset shows the highest proportion (46.6%) as "married-civ-spouse" (denotes a civilian spouse). They are the majority contributors to the group having incomes over $50k.

We also see for the "Never-married" group, a vast majority of the individuals earn less than or equal to. We could propose that they are not yet financially capable to support the obligations that come with having a family and as such wait longer to get married.

Also the remaining categories - Divorced, Married-AF-spouse, Married-spouse-absent, Separated and Widowed, are observed to majorly belong to the less than or equal to $50K income. A higher divorce-rate or separation hampers economic growth, as it increases the number of households, which requires more power and resources.



The percentage breakdown is as below:

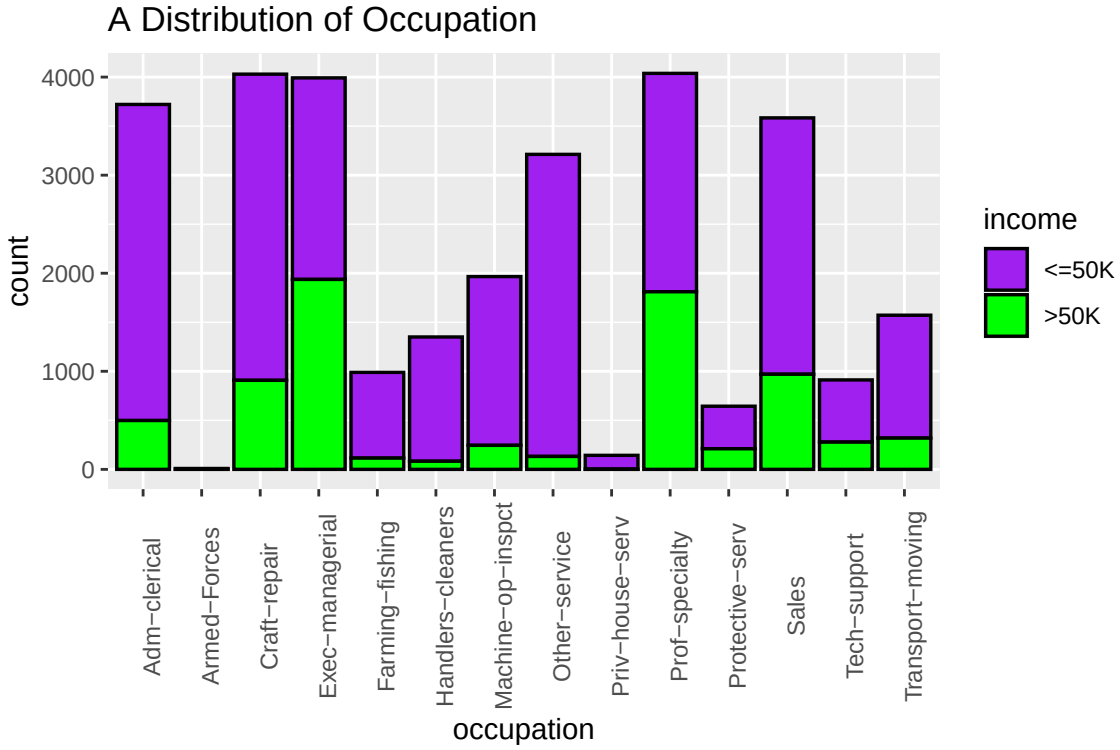
```
## # A tibble: 7 x 3
##   marital.status            n percent_marital_status
##   <chr>                 <int>                   <dbl>
## 1 Married-civ-spouse    14065                   46.6
## 2 Never-married          9726                   32.2
## 3 Divorced               4214                   14.0
## 4 Separated               939                    3.11
## 5 Widowed                 827                    2.74
## 6 Married-spouse-absent   370                    1.23
## 7 Married-AF-spouse        21                    0.0696
```

7) Occupation A study of the profession of the correspondents in the "Exec-managerial" and "Prof-specialty" groups shows there are almost equal proportion of individuals earning greater than $50K and less than or equal to $50K. These appear to be generally higher paying occupations in the corporate which require higher literacy to carry out critical tasks efficiently.

Although "Adm-clerical", "Other-service", "Craft-repair" and "Sales" being skilled occupations have a high proportion of wage-earners, they are primarily earning less than or equal to $50K with a smaller percentage of the population earning over $50K.

As anticipated occupations like “Farming-Fishing”, “Handlers-Cleaners”, “Machine-op-inspct”, “Protectiv-serv”, “Tech-support” and “Transport-moving” requiring lesser educational qualifications and are mainly labour-intensive vastly belong to the lower income bracket of less than or equal to $50K. We observe that government intervention would the number of jobs in the industries and also raise the average income of the employees.

A very minute fraction of the population (less than 0.05%) belong to niche professions like the “Armed-Forces” and “Priv-house-serv” and these are dominated by less than or equal to $50K earners.



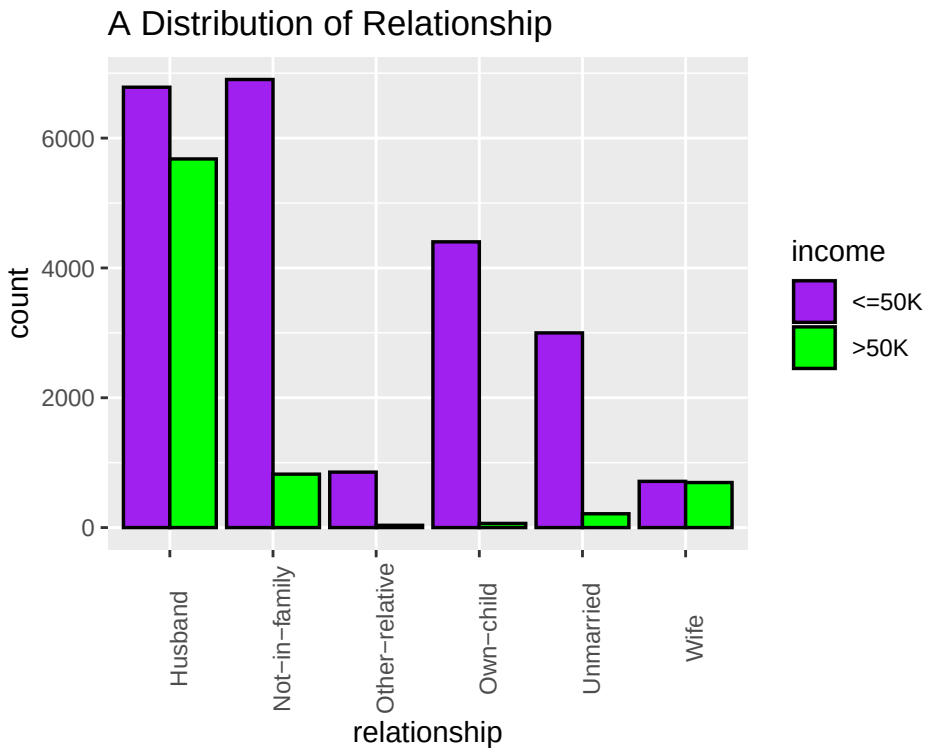
The percentage breakdown is as below:

```
## # A tibble: 14 x 3
##    occupation            n percent_occupation
##    <chr>             <int>              <dbl>
##  1 Prof-specialty     4038              13.4
##  2 Craft-repair       4030              13.4
##  3 Exec-managerial    3992              13.2
##  4 Adm-clerical       3721              12.3
##  5 Sales              3584              11.9
##  6 Other-service      3212              10.6
##  7 Machine-op-inspct  1966               6.52
##  8 Transport-moving   1572               5.21
##  9 Handlers-cleaners  1350               4.48
## 10 Farming-fishing     989               3.28
## 11 Tech-support        912               3.02
## 12 Protective-serv     644               2.14
## 13 Priv-house-serv     143               0.474
## 14 Armed-Forces          9               0.0298
```

8) Relationship A study of the relationship attribute shows that "husbands" majorly earn incomes over $50K in comparison to other members. We are aware that "husbands" are the primary or sole income earners in the household and generally cover most household expenses and financially support their dependents.

We see that "wives" are a small fraction (4.66%) of the employed population and there is equal distribution of greater than or less than or equal to $50K. In recent times, many wives are now leaving their traditional roles and going for higher education and taking on jobs that were once a monopoly of men.

Other members - "Not-in-family", "Other-relative", "Own-child" and "Unmarried" primarily belong to the less than or equal to $50K income bracket.

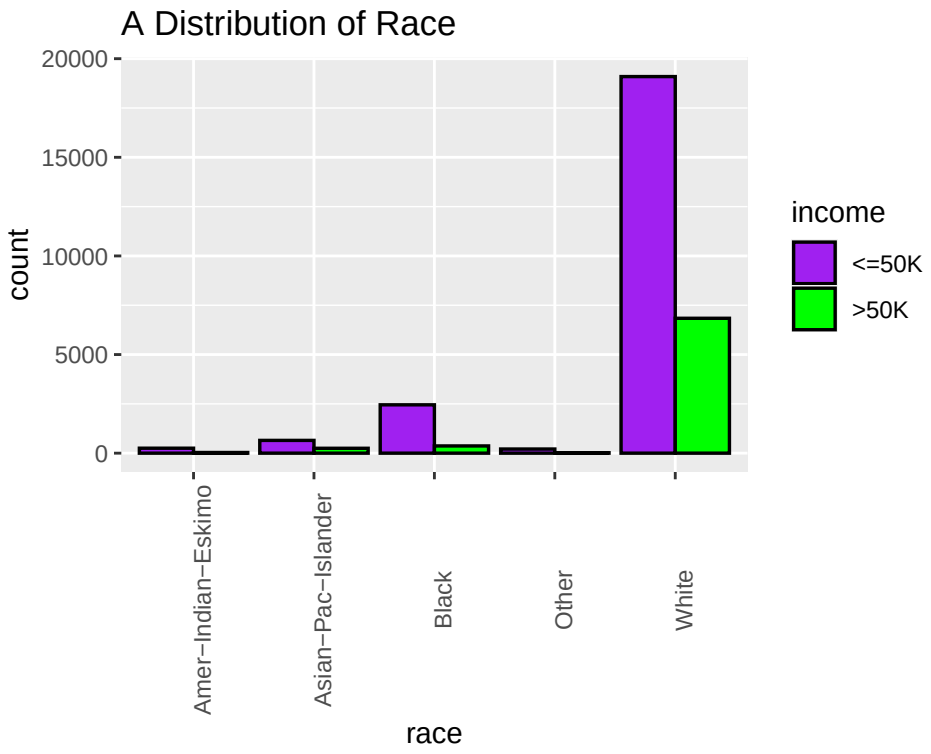


The percentage breakdown is as below:

```
## # A tibble: 6 x 3
##   relationship       n percent_relationship
##   <chr>          <int>                <dbl>
## 1 Husband        12463                41.3
## 2 Not-in-family   7726                25.6
## 3 Own-child       4466                14.8
## 4 Unmarried       3212                10.6
## 5 Wife            1406                 4.66
## 6 Other-relative   889                 2.95
```

9) Race A significant proportion (86%) of the race population are "white" and around a third of them earn an income of over $50K.

Among the others race, there is a small percentage of "black" (9.34%) and they primarily have an income of less than or equal to $50K. The remaining proportion of around 4% are of "Asian-Pac-Islander", "Amer-Indian-Eskimo" and "Other" races.



The percentage breakdown is as below:

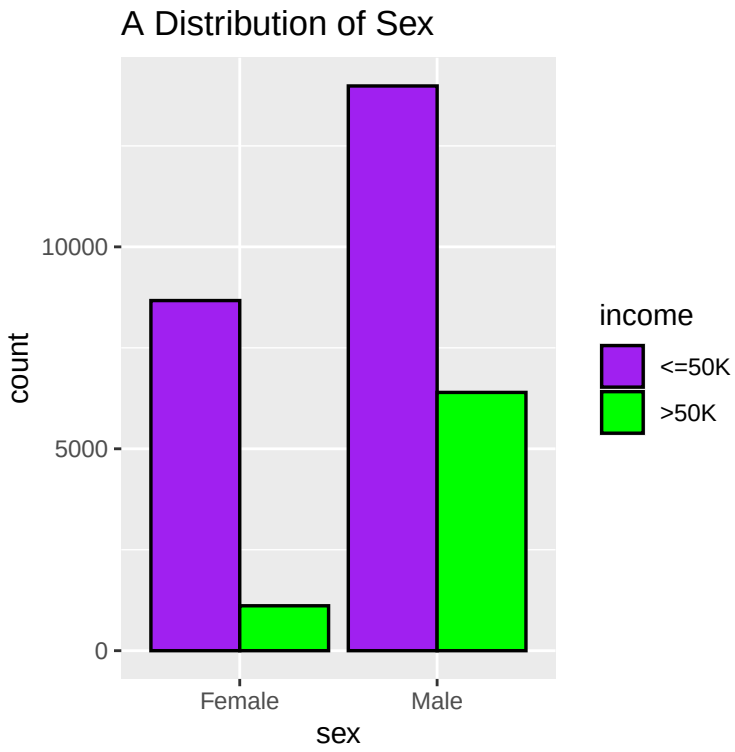
```
## # A tibble: 5 x 3
##   race                  n percent_race
##   <chr>             <int>        <dbl>
## 1 White             25933       86.0
## 2 Black              2817        9.34
## 3 Asian-Pac-Islander  895        2.97
## 4 Amer-Indian-Eskimo  286        0.948
## 5 Other               231        0.766
```

10) Sex The data set contains a higher proportion (67.6%) of males than females (32.4%). We also observe that proportion of males earning an income of greater than $50k is drastically higher than females earning more than %50K.

We are aware within high-paying occupations, women tend to be employed at lower levels of the occupational hierarchy while more men occupy senior positions. Around the world, occupations like teachers pay less than occupations like engineers. So gender differences in occupational choice affect gender differences in earnings.

If we study the earnings by industry or sector of economic activity, men are more likely to hold jobs at any skill level in manufacturing, a sector that pays relatively high earnings, while women are more likely to hold jobs in educational services, a sector that pays considerably less than manufacturing.

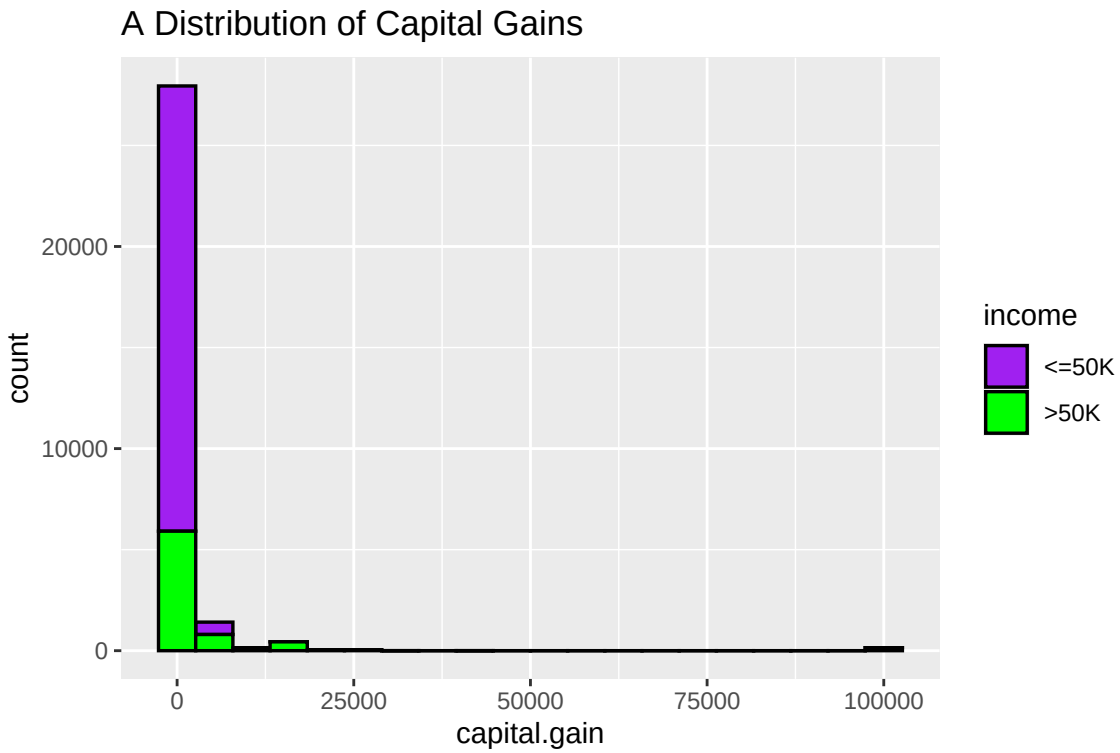
Woman are also more likely than men to work part-time due to family and household responsibilities.



The percentage breakdown is as below:

```
## # A tibble: 2 x 3
##   sex        n percent_sex
##   <chr>  <int>       <dbl>
## 1 Male   20380        67.6
## 2 Female  9782        32.4
```

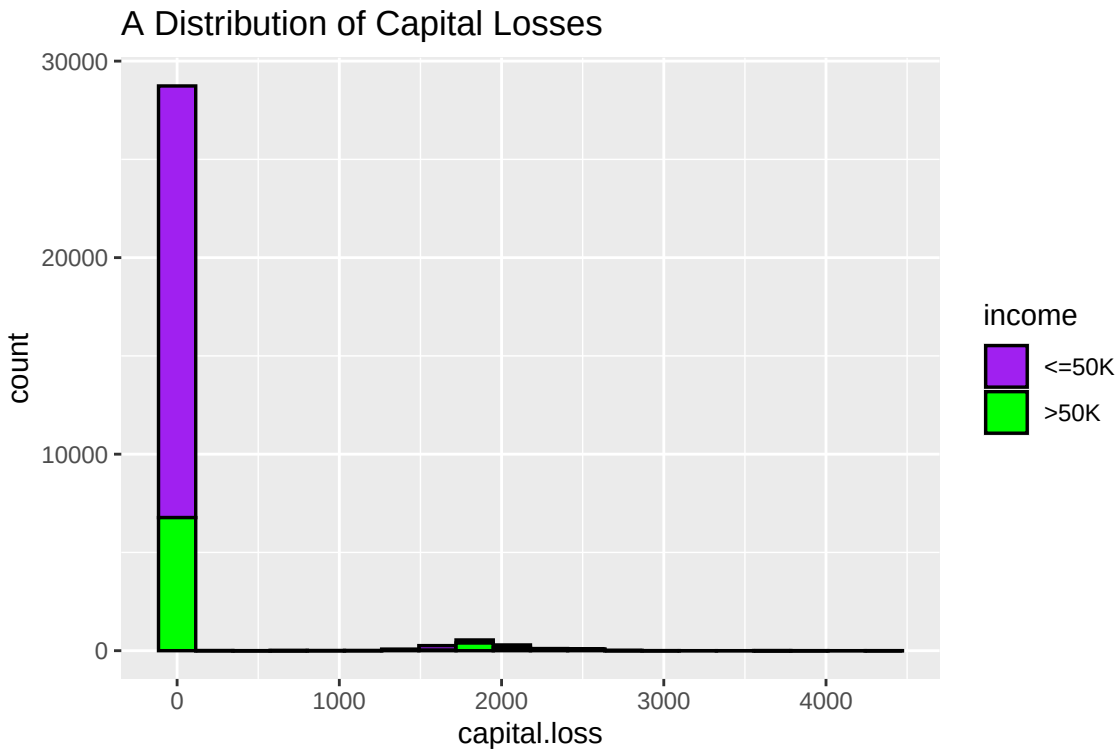
11) Capital gain The vast majority (91.6%) of the correspondents have zero capital gains and a majority of them earn an income of less than or equal to $50K.



The percentage breakdown is as below:

```
## # A tibble: 118 x 3
##    capital.gain     n percent_capital_gain
##           <int> <int>                <dbl>
##  1            0 27624                91.6
##  2        15024   337                 1.12
##  3         7688   270                 0.895
##  4         7298   240                 0.796
##  5        99999   148                 0.491
##  6         3103    94                 0.312
##  7         5178    91                 0.302
##  8         5013    69                 0.229
##  9         4386    67                 0.222
## 10         3325    53                 0.176
## # ... with 108 more rows
```

12) Capital loss The vast majority (95.3%) of the correspondents have zero capital losses and a majority of them earn an income of less than or equal to $50K.



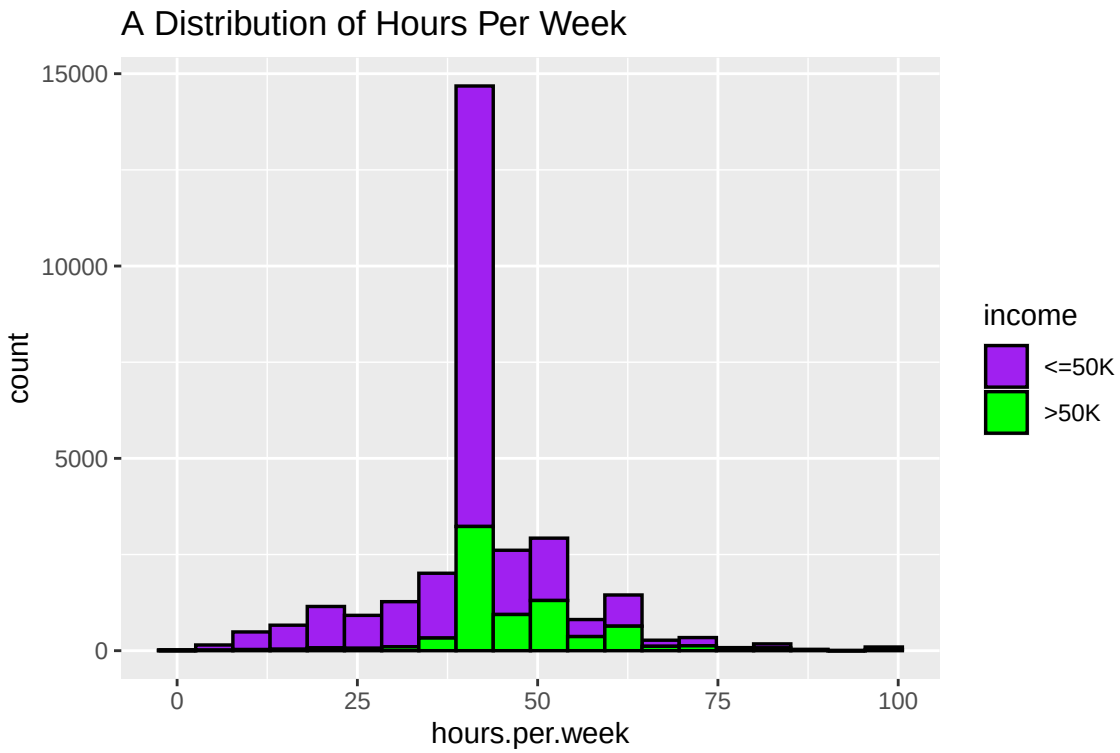
The percentage breakdown is as below:

```
## # A tibble: 90 x 3
##    capital.loss     n percent_capital_loss
##           <int> <int>                <dbl>
##  1            0 28735                95.3
##  2         1902   194                 0.643
##  3         1977   162                 0.537
##  4         1887   155                 0.514
##  5         1848    50                 0.166
##  6         1485    45                 0.149
##  7         2415    45                 0.149
##  8         1740    41                 0.136
##  9         1876    39                 0.129
## 10         1590    37                 0.123
## # ... with 80 more rows
```

13) Hours per week A study of the number of hours worked per week shows around half of the population have a standard 40 hour working week. The individuals working 40 hours per week tend to have a higher income of over $50K in comparison to those working less or more hours.

We also observe the a small population of individuals work more hours than the standard work week and a near equal distribution of them earn greater than $50K and less than or equal to $50K.

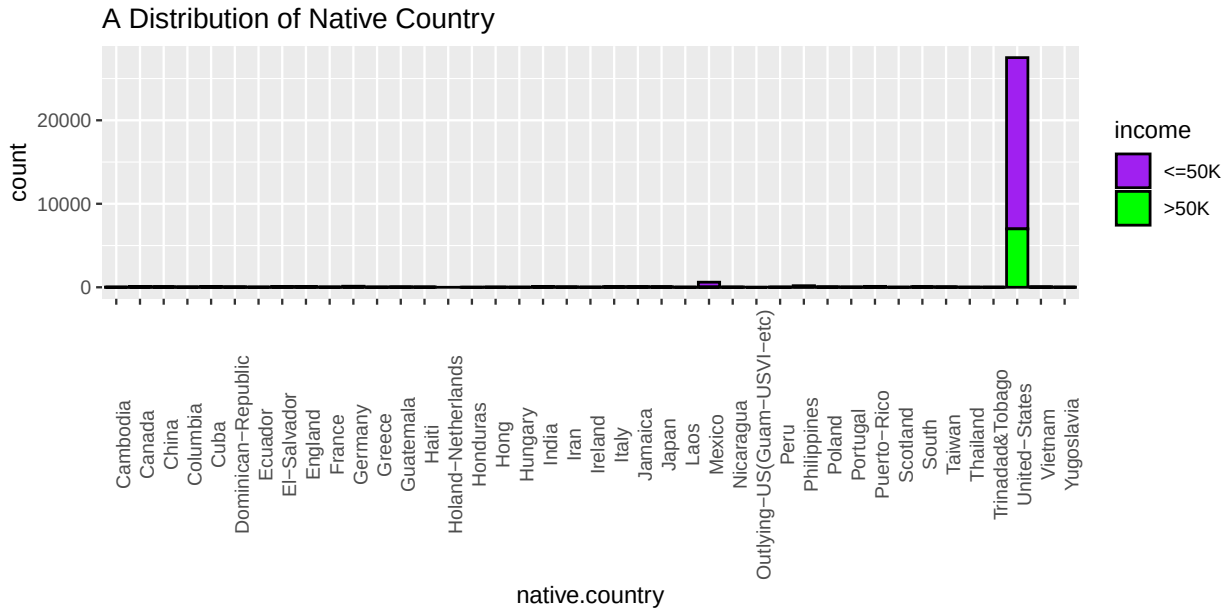
As expected, there is an extremely low proportion of those working less than 40 hours per week and earning over $50K.



The percentage breakdown is as below:

```
## # A tibble: 94 x 3
##    hours.per.week     n percent_hours_per_week
##             <int> <int>                   <dbl>
##  1             40 14251                   47.2
##  2             50  2718                    9.01
##  3             45  1753                    5.81
##  4             60  1405                    4.66
##  5             35  1184                    3.93
##  6             20  1054                    3.49
##  7             30   989                    3.28
##  8             55   672                    2.23
##  9             25   574                    1.90
## 10             48   494                    1.64
## # ... with 84 more rows
```

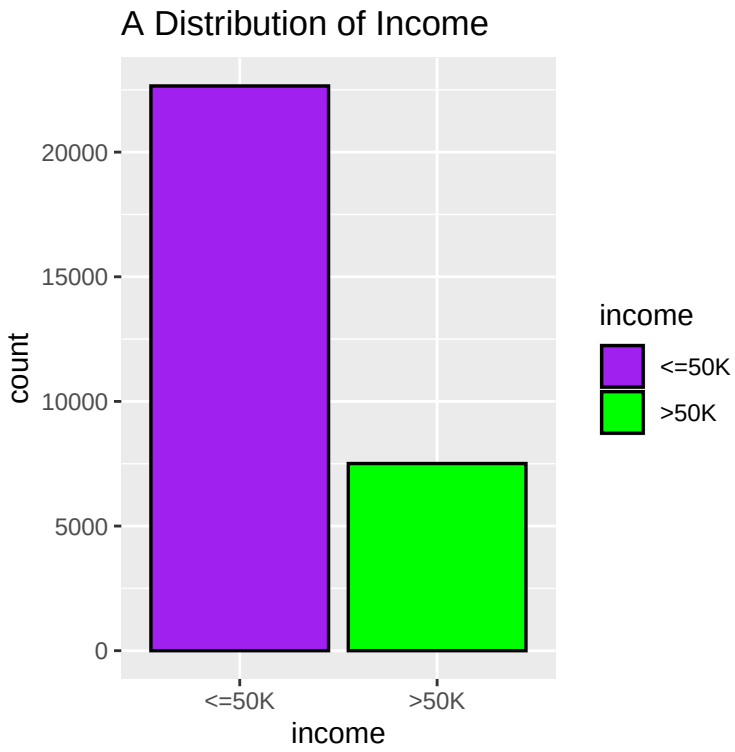
14) Native Country The vast majority (91.2%) of the correspondents are born in the United States and a majority of them earn an income of less than or equal to $50K.



The percentage breakdown is as below:

```
## # A tibble: 41 x 3
##    native.country     n percent_native_country
##    <chr>          <int>                  <dbl>
##  1 United-States  27504                  91.2
##  2 Mexico           610                   2.02
##  3 Philippines      188                   0.623
##  4 Germany          128                   0.424
##  5 Puerto-Rico      109                   0.361
##  6 Canada           107                   0.355
##  7 El-Salvador      100                   0.332
##  8 India            100                   0.332
##  9 Cuba              92                   0.305
## 10 England           86                   0.285
## # ... with 31 more rows
```

15) Income Around 3/4th of the population in the "Adult Census Income" data set earn an income of less than or equal to $50K and the remaining 1/4th earn greater than $50K annually as seen below.



The percentage breakdown is as below:

```
## # A tibble: 2 x 3
##   income     n percent_income
##   <chr>  <int>          <dbl>
## 1 <=50K  22654           75.1
## 2 >50K    7508           24.9
```

In the data analysis, records classified as greater than $50 thousand (>50K) are assigned a value of 1 and the rest (<=50K) are assigned 0. In the Regression Tree analysis the categorical values are used in addition to the numerically assigned ones.

2.2.2 Data Visualization

We will convert the predictor "income" to a factor with 2 levels- less than or equal to $50K and greater than $50K.

```
adult_census_income_dataset$income <- as.factor(adult_census_income_dataset$income)
class(adult_census_income_dataset$income)
```

```
## [1] "factor"
```

```
str(adult_census_income_dataset)
```

```
## 'data.frame':    30162 obs. of  15 variables:
##  $ age           : int  82 54 41 34 38 74 68 45 38 52 ...
##  $ workclass     : chr  "Private" "Private" "Private" "Private" ...
```

```
##  $ fnlwgt        : int  132870 140359 264663 216864 150601 88638 422013 172274 164526 129177 ...
##  $ education     : chr  "HS-grad" "7th-8th" "Some-college" "HS-grad" ...
##  $ education.num : int  9 4 10 9 6 16 9 16 15 13 ...
##  $ marital.status: chr  "Widowed" "Divorced" "Separated" "Divorced" ...
##  $ occupation    : chr  "Exec-managerial" "Machine-op-inspct" "Prof-specialty" "Other-service" ...
##  $ relationship  : chr  "Not-in-family" "Unmarried" "Own-child" "Unmarried" ...
##  $ race          : chr  "White" "White" "White" "White" ...
##  $ sex           : chr  "Female" "Female" "Female" "Female" ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 3900 3900 3770 3770 3683 3683 3004 2824 2824 ...
##  $ hours.per.week: int  18 40 40 45 40 20 40 35 45 20 ...
##  $ native.country: chr  "United-States" "United-States" "United-States" "United-States" ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 2 1 2 2 2 ...
```

2.2.3 Data Partioning

a) Split the "Adult Census Income" dataset into train and test (validation) sets.

NOTE: The validation data will NOT be used for training the algorithm and will ONLY be used for evaluating the accuracy of the final algorithm.

```
set.seed(1)

test_index <- createDataPartition(adult_census_income_dataset$income, times = 1, p = 0.2, list = FALSE)
adult_census_income_training<- adult_census_income_dataset[-test_index, ]
adult_census_income_validation <- adult_census_income_dataset[test_index, ]
```

b) Split the "Adult Census Income" train dataset into train and test sets to train our algorithms

```
set.seed(10)
test_index1 <- createDataPartition(adult_census_income_training$income, times = 1, p = 0.2, list = FALSE
train_set <- adult_census_income_training[-test_index1, ]
test_set <- adult_census_income_training[test_index1, ]
```

2.3 Modelling Approach

2.3.1 Terminology

a) Model: A machine learning model can be a mathematical representation of a real-world process. To generate a machine learning model we need to provide training data to a machine learning algorithm to learn from.

b) Algorithm: Machine Learning algorithm is the hypothesis set that is taken at the beginning before the training starts with real-world data. For example, when we say Linear Regression algorithm, it means a set of functions that define similar characteristics as defined by Linear Regression and from those set of functions we will choose one function that fits the most by the training data.

c) Training: While training for machine learning, you pass an algorithm with training data. The learning algorithm finds patterns in the training data such that the input parameters correspond to the outcome. The output of the training process is a machine learning model which you can then use to make predictions. This process is also called "learning".

d) Regression: Regression techniques are used when the output is real-valued based on continuous variables. For example, any time series data. This technique involves fitting a line.

e) Classification: In classification, you will need to categorize data into predefined classes. For example, an email can either be 'spam' or 'not spam'.

f) Outcome: The outcome is whatever the output of the input variables. It could be the individual classes that the input variables maybe mapped to in case of a classification problem or the output value range in a regression problem. If the training set is considered then the outcome is the training output values that will be considered.

g) Feature: Features are individual independent variables that act as the input in your system. Prediction models use features to make predictions.

h) Overfitting: An important consideration in machine learning is how well the approximation of the outcome function that has been trained using training data, generalizes to new independent data. Generalization works best if the signal or the sample that is used as the training data has a high signal to noise ratio. If that is not the case, generalization would be poor and we will not get good predictions. A model is overfitting if it fits the training data too well and there is a poor generalization of new data.

i) Regularization: Regularization is the method to estimate a preferred complexity of the machine learning model so that the model generalizes and the over-fit/under-fit problem is avoided. This is done by adding a penalty on the different parameters of the model thereby reducing the freedom of the model.

j) Parameter: Parameters are configuration variables that can be thought to be internal to the model as they can be estimated from the training data. Algorithms have mechanisms to optimize parameters. These are some key machine learning terms that I thought are important and should be looked into when studying this project.

2.3.2 Accuracy

Machine learning model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a data set based on the input, training or data. The better a model can generalize to 'unseen' data, the better predictions and insights it can produce, which in turn deliver more business value.

Formally, accuracy has the following definition:

$$Accuracy = \frac{Numberofcorrectpredictions}{Totalnumberofpredictions}$$

Companies use machine learning models to make practical business decisions, and more accurate model outcomes result in better decisions. The cost of errors can be huge, but optimizing model accuracy mitigates that cost.

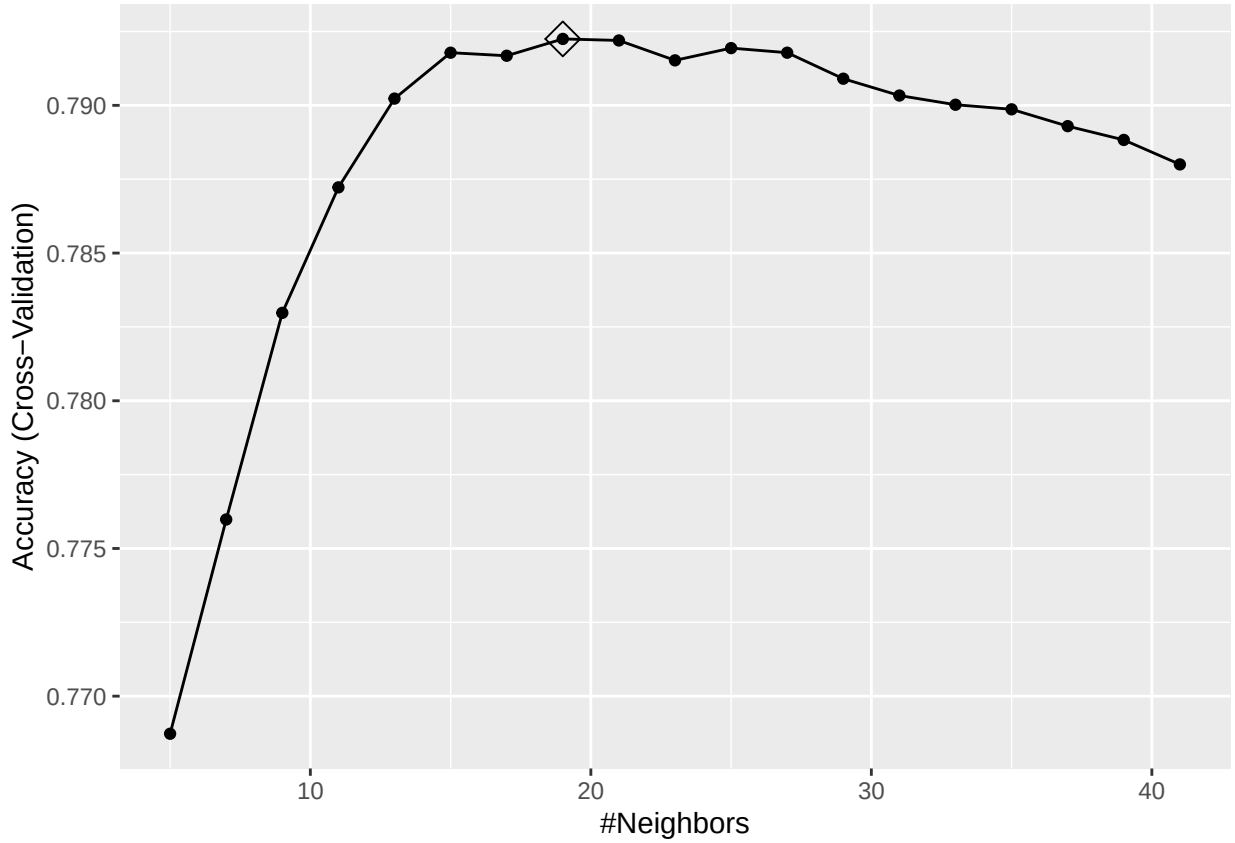
2.4 Models

I. K Nearest Neighbours (KNN) Model We will apply the cross-validation method built into the caret package on the k-value ranging from 4 to 41. We will use a 10-fold cross-validation to reduce the time taken to run and to avoid over-fitting.

We use k as a tuning parameter which represents the number of neighbors to be considered. We will use a k-value ranging from 4 to 41.

```r
#Train the knn model on the training data set
set.seed(9)
#Use a 10 fold cross-validation method
control <- trainControl(method = "cv", number = 10, p = .9)
train_knn <- train(income ~ .,
                   method = "knn",
                   data = train_set,
                   tuneGrid = data.frame(k = seq(5,41,2)),
                   trControl = control)

#Plot the k values
ggplot(train_knn, highlight = TRUE)
```




```r
#Choose the optimal k value
train_knn$bestTune
```

```
##    k
## 8 19
```

```r
#Compute the accuracy of the knn model on the test data set
knn_accuracy <- confusionMatrix(predict(train_knn, test_set, type = "raw"),
                                test_set$income)$overall["Accuracy"]

#Create a results table to store the results for each model
```

```r
accuracy_results <- bind_rows(data.frame(method = "KNN Model", Accuracy = knn_accuracy))

#View the knn accuracy results in the table
accuracy_results %>% knitr::kable()
```

method	Accuracy
KNN Model	0.7872385

We obtain an accuracy of 78.72% on the test set.

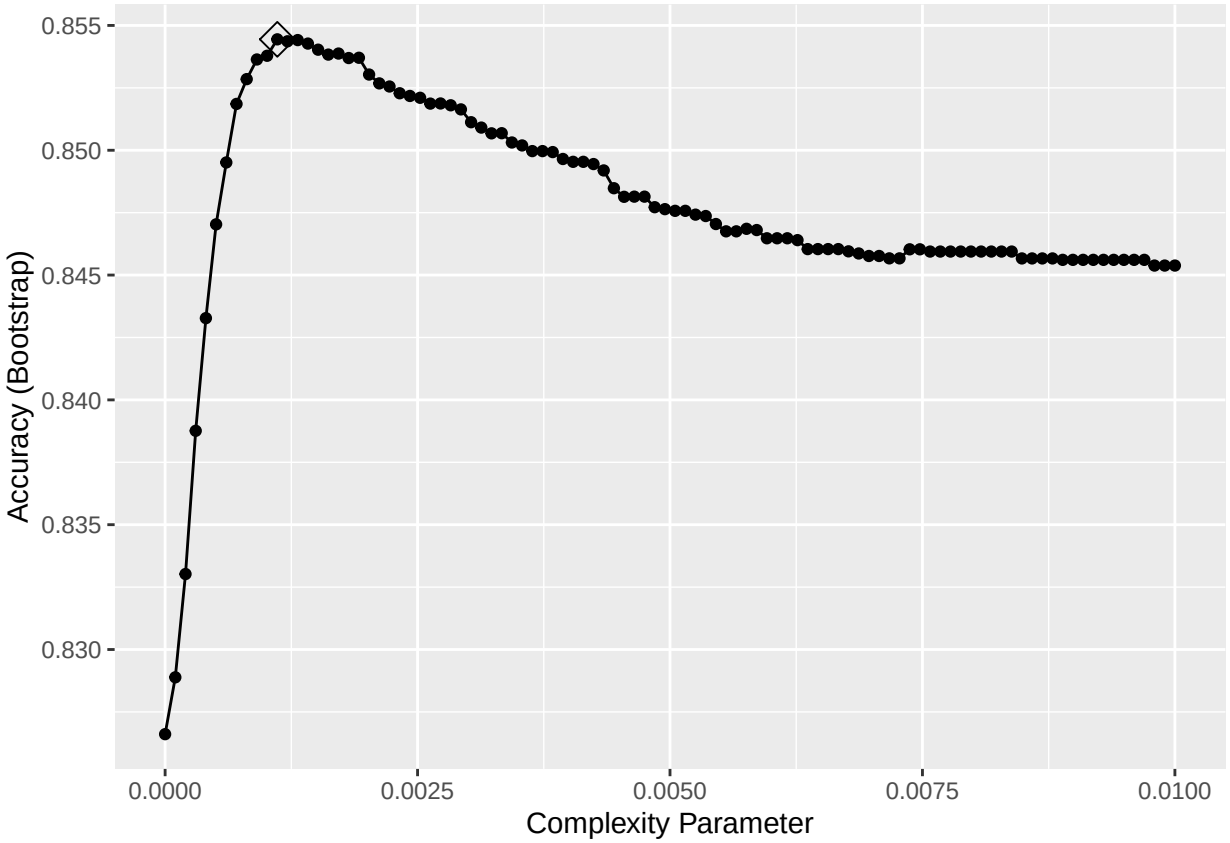
II. Classification and Regression Trees (CART) Model We will train a CART algorithm using the "rpart" method from the caret package. We will use cross-validation to choose the best cp (complexity parameter).

RPART (Recursive Partitioning And Regression Trees) complexity measure is a combination of the size of a tree and the ability of the tree to separate the classes of the target variable. If the next best split in growing a tree does not reduce the tree's overall complexity by a certain amount, rpart will terminate the growing process.

A tree is similar to a flow chart with yes or no questions and predictions made at the ends that are called nodes. Decision trees are a type of supervised learning algorithm that work by partitioning the predictor space in order to predict an outcome (the "income" in our case). The partitions are created recursively.

```r
#Train the CART model using rpart on training set
set.seed(300)
train_rpart <- train(income ~ .,
                     method = "rpart",
                     tuneGrid = data.frame(cp = seq(0, 0.01, len=100)),
                     data = train_set)

#Highlight the optimized complexity parameter
ggplot(train_rpart, highlight=TRUE)
```




```
#Obtain optimal cp
train_rpart$bestTune
```

```
##             cp
## 12 0.001111111
```

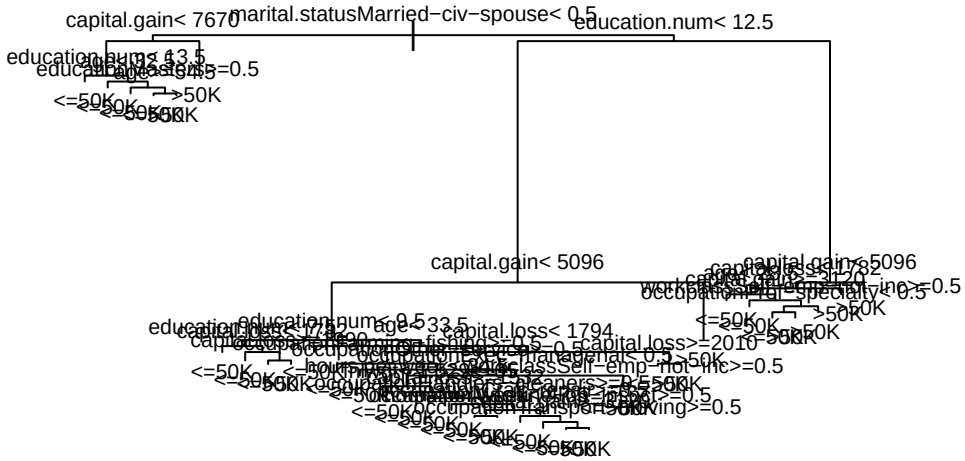
```
#Compute the accuracy of the CART model on the test data set
rpart_accuracy <- confusionMatrix(predict(train_rpart, test_set),
                                  test_set$income)$overall["Accuracy"]

#Store the results of the model
accuracy_results <- bind_rows(accuracy_results,data.frame(method="CART Model", Accuracy = rpart_accuracy

#View the CART accuracy results in the table
accuracy_results %>% knitr::kable()
```

method	Accuracy
KNN Model	0.7872385
CART Model	0.8479387

```
#Classification tree figure
plot(train_rpart$finalModel, margin = 0.1)
text(train_rpart$finalModel, cex = 0.7)
```



We obtain an accuracy of 84.79% which is higher than the KNN model.

III. Gradient Boosting Machines (GBM) Model We will train a GBM algorithm using the "gbm" method from the caret package. We will use a 10-fold cross-validation to reduce the time taken to run and to avoid over-fitting.

```
#Train the gbm model on the training data set
set.seed(2000)

#Use a 10 fold cross-validation method
trCtrl <- trainControl (method = "cv", number = 10)

#Train the gbm model
train_gbm <- train (income~ .,
                    trControl = trCtrl,
                    method = "gbm",
                    preProc="zv",
                    data = train_set,
                    verbose = FALSE)

#Compute the accuracy of the gbm model on the test data set
gbm_accuracy <- confusionMatrix(predict(train_gbm, test_set, type = "raw"),
                                test_set$income) $overall["Accuracy"]
gbm_accuracy
```

```
##  Accuracy
## 0.8560182
```

```
#Store the results of the model
accuracy_results <- bind_rows(accuracy_results,data.frame(method="GBM Model",Accuracy = gbm_accuracy))

#View the gbm accuracy results in the table
accuracy_results %>% knitr:: kable()
```

method	Accuracy
KNN Model	0.7872385
CART Model	0.8479387
GBM Model	0.8560182

We obtain an accuracy of 85.6% which is higher than the CART model.

IV. Random Forest (RF) Model We will train a RF algorithm using the "randomForest" package in R.

```
#Train the rf model on the training data set
set.seed(9)
train_rf <- randomForest(income ~ ., data = train_set)

#Compute the accuracy of the rf model on the test dataset
rf_accuracy <- confusionMatrix(predict(train_rf, test_set),
                               test_set$income)$overall["Accuracy"]

#Store the results of the model
accuracy_results <- bind_rows(accuracy_results,data.frame(method="RF Model", Accuracy = rf_accuracy))

#View the rf accuracy results in the table
accuracy_results %>% knitr::kable()
```

method	Accuracy
KNN Model	0.7872385
CART Model	0.8479387
GBM Model	0.8560182
RF Model	0.8564326

The accuracy is greatly improved to 85.64% which is higher in comparison to all previous models.

While we cannot plot the tree in the same way as for the classification tree above, it is still possible to see the importance of each variable, measured by the Mean Decrease in Gini.

```
#View the rf importance
importance(train_rf)
```

```
##                 MeanDecreaseGini
```

```
## age                  721.79117
## workclass            214.66492
## fnlwgt               637.04360
## education            220.68866
## education.num        696.36492
## marital.status       499.60426
## occupation           355.91625
## relationship         740.34162
## race                  77.52547
## sex                   94.70661
## capital.gain         802.41878
## capital.loss         216.00225
## hours.per.week       440.88980
## native.country        90.09243
```

2.5 Validation

From our results table, we can see that the Random Forest model achieved the highest accuracy of 85.64%. We will use the algorithm trained and tested above to now test the model on our validation set.

We will use the entire "adult_census_income_training" training set to predict the income results on the "adult_census_income_validation" validation set.

```
set.seed(3)
final_train_rf <- randomForest(income ~ ., data = adult_census_income_training)

#Compute the accuracy of the rf model on the validation data set
final_rf_accuracy <- confusionMatrix(predict(final_train_rf, adult_census_income_validation),
                                     adult_census_income_validation$income)$overall["Accuracy"]

#Store the results of the model
accuracy_results_val <- (data.frame(method="Random Forest Model", Accuracy = final_rf_accuracy))

##View the rf accuracy results in the table
accuracy_results_val %>% knitr::kable()
```

	method	Accuracy
Accuracy	Random Forest Model	0.8523123

We obtain a similar accuracy as achieved during the testing phase.

The final model achieves 85.23% accuracy on the validation data set.

3. Results

As observed in the results table, the Random Forest Model has the highest overall accuracy of 85.23% for predicting whether a person's income is less than or equal to $50K or greater than $50K in the "Adult Census Income" data set.

```r
#Results
accuracy_results_val %>% knitr::kable()
```

	method	Accuracy
Accuracy	Random Forest Model	0.8523123

3.1 Discussion

The "Adult Census Income" data set is of 1994 census, which is of 26 years old, refers to the U.S population. Accordingly, the results would apply to around 5 years or maybe upto the next census (i.e 2004).

An important note to consider is that we have not removed any of the variables in the data set as we wanted to explore the nature of the data and gain realistic insights from our analysis. Accordingly, we can study some of the inferences from the drivers in the "Importance" table. A higher "Mean Decrease" in Gini indicates higher variable importance.

```r
#View the rf importance
importance(final_train_rf)
```

```
##                 MeanDecreaseGini
## age                     916.67698
## workclass               262.32493
## fnlwgt                  794.49973
## education               280.16845
## education.num           831.69440
## marital.status          634.98266
## occupation              432.21439
## relationship            920.68221
## race                     94.15127
## sex                     110.69378
## capital.gain            999.16934
## capital.loss            278.27539
## hours.per.week          543.67954
## native.country          109.14923
```

"Capital.gain" was one of the highest predictors which was expected as around 91.6% of the values were '0'. "Relationship" proved to be a strong predictor as observed during analysis that husbands are the primary earners of a household. Also "age" was a good predictor as we witnessed a large proportion of the population in the "Prime Working" age of 25 to 54 years.

The least important predictors as anticipated appeared to be "race" and "native.country" as these were generally skewed towards "White" and "US born citizens".

The main aspects are: around 2/3rd of the population is males, number of white people are more (86%) and fewer immigrants (around 9%). Accordingly, this must be factored in if using the algorithm to create predictions on current populations.

During the data analysis stage as seen above we have suggested various measures and interventions to be taken by the government.

We trained and tested a total of four algorithms in our project study of "Adult Census Income" data set. The Random Forest Model had a final accuracy of 85.23% for predicting whether a person's income is less than or equal to $50K or greater than $50K.

4. Conclusion

We first trained and tested the KNN Model and observed a fundamental weakness of it not being able to handle categorical features very efficiently.

Hence, we moved on to the CART model which improved the accuracy considerably to 84.79% from 78.72% obtained through the KNN Model. However, a big limitation is the fact that it is a non-parametric technique; it is not recommended to make any generalization on the underlying phenomenon based upon the results observed. Although the rules obtained through the analysis can be tested on new data, it must be remembered that the model is built based upon the sample without making any inference about the underlying probability distribution. In addition to this, another limitation of CART is that the tree becomes quite complex after seven or eight layers. Interpreting the results in this situation is not intuitive.

We then progressed to the GBM model. There was a slight increase in the accuracy to 85.6%. It works well with categorical variables and provides unmatched predictive accuracy. It is flexible and can optimize on different loss functions and provides several hyper-parameter tuning options that make the function fit very flexible. However, GBM will continue improving to minimize all errors. This can overemphasize outliers and cause over-fitting. To neutralize this effect we used cross-validation. It often requires many trees (>1000) which can be time and memory exhaustive.

The final model we trained was Random Forests in which we achieved an improved accuracy of 85.64%. We found Random Forest to be a flexible, easy to use machine learning algorithm that produced, even without hyper-parameter tuning, a predictive result in our project study. It can be used for both classification and regression tasks.

Random forests provided the highest accuracy of 85.23% on the validation set. The random forest technique can also handle big data with numerous variables running into thousands. It has the power to handle a large data set with higher dimensionality.

4.1 Future Work

One of the main limitations of the data set is that it is over 26 years old and is not a realistic representation of the current US Population and demographics. As such predictions made on present census data would perform quite differently.

As so many of the variables influencing income have changed in present times it would be imperative to train new machine learning models on more recent census data. Cutoff income of $50K would not be relevant here and would need to be scaled upwards towards the current trend of income. Also migrant population would have increased as many citizens from foreign countries would have migrated to the U.S for better prospects. We can look at additional variables that would affect the income of an individual like family size, state or county where the individual is located, as well as numerical variables like area-wise cost of living differentials and so on.