

HarvardX PH125.9x - Capstone Project - CYO

Predicting Income Using Knn and randomForest

Kwaku Owusu-Tieku

2/9/2022

Introduction

Income disparity is common in every society and it's probably been around since man has been able to measure it. There are various reasons why income disparities exist. These include age, gender, race, level of education, etc.

For instance, with respect to age, younger people who have just entered the workforce may earn less than their counterparts who have been in the industry for many years. The level of education may also explain income disparities since some high paying jobs often require skilled and highly educated individuals.

Even different geographic regions or cities may offer different levels of income since some high paying jobs are sometimes more concentrated in some regions or cities than others. However, it is still difficult to measure how much these factors determine one's income.

In this project, I will I apply kNN and randomForest to a dataset containing various attributes (predictors) and individual's income level (above or below \$50K), and predict the individual's income level.

The kNN and randomForest algorithms were chosen because they are very easy to train, and they are able handle a wide range data including a mix of categorical and numeric variables and offer good performance.

The original data

Source

The data used in this project is part of the UCI machine learning datasets provided to the public. It can be found at: <https://archive.ics.uci.edu/ml/machine-learning-databases/adult/>.

Three main files are provided as part of the dataset.

- `adult.names`
- `adult.data`
- `adult.test`

The `adult.names` file, provides very useful information about the data. The `adult.data` and `adult.test` files are the training and test sets. The data had already been split into training and test sets by the contributor(s). The training data contains *32561* records. The test set contains *16281* records, about 50% of the training set. This provides a good amount of data for training while leaving enough for validation. Each record in the datasets represents an observation.

The documentation indicates that the original data was retrieved from the 1994 census bureau database (<http://www.census.gov/ftp/pub/DES/www/welcome.html>), so the samples were taken from real dataset. The U.S. population and workforce have changed dramatically since 1994, so the patterns in this data may be different from when the samples were originally taken.

Reading the data

A quick peek at the data reveals that there are no column names. However, descriptions in the `adult.names` file provides useful information in naming the columns.

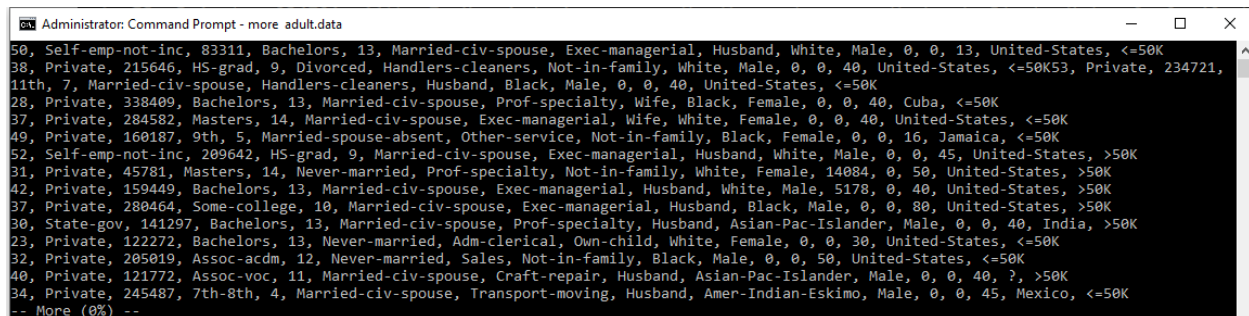


Figure 1: original data - raw

Code to read the data

```
# use wrapr::qc() function to create a list of quoted string for the column headers
COL_HEADERS <- qc(age, workclass, final_weight, education, education_num,
  marital_status, occupation, relationship, race, sex,
  capital_gain, capital_loss, hours_per_week, native_country,
  Income)

# read training and test datasets into memory
adult_income_train <- read.csv(TRAINING_DATA_PATH, col.names = COL_HEADERS)
adult_income_validation <- read.csv(VALIDATION_DATA_PATH, col.names = COL_HEADERS)

# trim all character columns
adult_income_train <- adult_income_train %>% mutate(across (where(is.character), str_trim))
adult_income_validation <- adult_income_validation %>% mutate(across (where(is.character), str_trim))
```

Structure of the dataframe

The data contains 15 columns; the *Income* column is the response variable with the other fourteen columns being the independent variables. The structure of the dataframe looks as below:

```
glimpse(adult_income_train)
```

Rows: 32,560

Columns: 15

```
$ age          <int> 50, 38, 53, 28, 37, 49, 52, 31, 42, 37, 30, 23, 32, 40, ~
$ workclass    <chr> "Self-emp-not-inc", "Private", "Private", "Private", "P~
$ final_weight <int> 83311, 215646, 234721, 338409, 284582, 160187, 209642, ~
$ education    <chr> "Bachelors", "HS-grad", "11th", "Bachelors", "Masters", ~
$ education_num <int> 13, 9, 7, 13, 14, 5, 9, 14, 13, 10, 13, 13, 12, 11, 4, ~
$ marital_status <chr> "Married-civ-spouse", "Divorced", "Married-civ-spouse", ~
$ occupation   <chr> "Exec-managerial", "Handlers-cleaners", "Handlers-clean~
$ relationship <chr> "Husband", "Not-in-family", "Husband", "Wife", "Wife", ~
$ race         <chr> "White", "White", "Black", "Black", "White", "Black", "~
$ sex          <chr> "Male", "Male", "Male", "Female", "Female", "Female", "~
$ capital_gain <int> 0, 0, 0, 0, 0, 0, 0, 14084, 5178, 0, 0, 0, 0, 0, 0, ~
$ capital_loss <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ hours_per_week <int> 13, 40, 40, 40, 40, 16, 45, 50, 40, 80, 40, 30, 50, 40, ~
$ native_country <chr> "United-States", "United-States", "United-States", "Cub~
$ Income       <chr> "<=50K", "<=50K", "<=50K", "<=50K", "<=50K", "<=50K", "~
```

Data quality checks

Although the data for this analysis has previously been cleaned by the team that made it available to UCI, a few things had to be checked to make sure the data is ready to be transformed and used in an analysis.

- 1. Verify correct number of records read.** Based on the documentation in the *adult.names* documentation, there are 32561 records in the training set.
- 2. Check for invalid/missing data (NAs, invalid zeros for age, etc).** Verify the presence of NA in all columns. No NAs (null values) are found in both train and test sets. However, some of the categorical columns have questions mark (?) as values. Again, according to the documentation, these were missing values; the NAs (null values) had been replaced with questions marks.
- 3. Numeric variables.** Check the range of values to make sure they do not include invalid values, e.g. age should be greater than zero, and not excessively high number.
- 4. Categorical values.** Check the number of levels.

The following code was used for the basic sanity checks described above.

Basic data quality checks - data read

```
#####  
## DATA QUALITY CHECKS  
#####  
  
# 1. check that correct number of rows was read  
TRAIN_NUM_OF_ROWS_EXPECTED <- 32561  
VALIDATION_NUM_OF_ROWS_EXPECTED <- 16281  
  
verify_stats <- data.frame("Item" = c("Number of row (training set)",  
                                     "Number of row (test set)"))  
verify_stats$Expected <- c(TRAIN_NUM_OF_ROWS_EXPECTED, VALIDATION_NUM_OF_ROWS_EXPECTED)  
verify_stats$Actual <- c(nrow(adult_income_train), nrow(adult_income_validation))  
  
#2. check for NA in all columns  
verify_stats$HasNA <- c(any(is.na(adult_income_train)), any(is.na(adult_income_validation)))  
  
colnames(verify_stats) <- c("", "Expected", "Actual", "Has Nulls")  
kbl(verify_stats, booktabs = T, caption = "Basic data quality checks") %>%  
kable_styling(latex_options = c("striped", "hold_position")) %>%  
  column_spec(3, color = "red") %>%  
  column_spec(4, color = "blue")
```

Table 1: Basic data quality checks

	Expected	Actual	Has Nulls
Number of row (training set)	32561	32560	FALSE
Number of row (test set)	16281	16280	FALSE

From the table, it appears that each dataset is off by one record. Since this is not a large number, I will ignore it. We also see that there are no NAs in each dataset.

Basic data quality check - min/max for numeric data

```
#3. numeric columns: check range of values, min/max

# function to check min/max
fnFindMinMax <- function(myFeature){
  data.frame("min" = min(myFeature), "max" = max(myFeature))
}

supply( adult_income_train %>% select(where(is.numeric)), fnFindMinMax) %>%
  kbl(booktabs=T, caption="Numeric Variables min/max") %>%
  kable_styling(latex_options = c("striped", "hold_position")) %>%
  column_spec(0, bold=T) %>%
  row_spec(0, bold=T)
```

Table 2: Numeric Variables min/max

	age	final_weight	education_num	capital_gain	capital_loss	hours_per_week
min	17	12285	1	0	0	1
max	90	1484705	16	99999	4356	99

Basic data quality check - categorical data unique values

```
#4. for categorical values, inspect the unique values
fnShowUniqueValues <- function(myFeature){
  unique(myFeature)
}

adult_income_train %>% select(where(is.character)) %>% supply(fnShowUniqueValues)
```

```
$workclass
[1] "Self-emp-not-inc" "Private"          "State-gov"        "Federal-gov"
[5] "Local-gov"        "?"                "Self-emp-inc"     "Without-pay"
[9] "Never-worked"

$education
[1] "Bachelors"      "HS-grad"      "11th"          "Masters"      "9th"
[6] "Some-college"   "Assoc-acdm"   "Assoc-voc"     "7th-8th"      "Doctorate"
[11] "Prof-school"    "5th-6th"      "10th"          "1st-4th"      "Preschool"
[16] "12th"

$marital_status
[1] "Married-civ-spouse" "Divorced"       "Married-spouse-absent"
[4] "Never-married"      "Separated"      "Married-AF-spouse"
[7] "Widowed"

$occupation
[1] "Exec-managerial" "Handlers-cleaners" "Prof-specialty"
```

```

[4] "Other-service"      "Adm-clerical"      "Sales"
[7] "Craft-repair"       "Transport-moving"   "Farming-fishing"
[10] "Machine-op-inspct"  "Tech-support"      "?"
[13] "Protective-serv"    "Armed-Forces"      "Priv-house-serv"

$relationship
[1] "Husband"          "Not-in-family"    "Wife"              "Own-child"
[5] "Unmarried"        "Other-relative"

$race
[1] "White"            "Black"            "Asian-Pac-Islander"
[4] "Amer-Indian-Eskimo" "Other"

$sex
[1] "Male"    "Female"

$native_country
[1] "United-States"      "Cuba"
[3] "Jamaica"            "India"
[5] "?"                  "Mexico"
[7] "South"              "Puerto-Rico"
[9] "Honduras"           "England"
[11] "Canada"             "Germany"
[13] "Iran"               "Philippines"
[15] "Italy"              "Poland"
[17] "Columbia"           "Cambodia"
[19] "Thailand"           "Ecuador"
[21] "Laos"               "Taiwan"
[23] "Haiti"              "Portugal"
[25] "Dominican-Republic" "El-Salvador"
[27] "France"             "Guatemala"
[29] "China"              "Japan"
[31] "Yugoslavia"         "Peru"
[33] "Outlying-US(Guam-USVI-etc)" "Scotland"
[35] "Trinidad&Tobago"    "Greece"
[37] "Nicaragua"          "Vietnam"
[39] "Hong"               "Ireland"
[41] "Hungary"            "Holand-Netherlands"

$Income
[1] "<=50K" ">50K"

```

From the listings, there are no unusual values for either the numeric or the categorical data. As mentioned earlier, three variables (workclass, occupation, and native_country) do have the “?” as one of the values, meaning these three variables originally had null values which have been replaced. All the data, including the “?”, will be used in the analysis.

Data Visualization and Exploration

Equal access is not equal gain

Even though one can argue that the U.S. population has equal access to education and employment, we know from experience that there are many factors or barriers that force people into certain income groups. These factors includes levels of education, gender, race, geographic areas people live in, to mention a few.

In this section, I explore the data to find general properties and patterns. The dataset contains many categorical features, so I am very interested in learning how income differs across different categories. Since the response variable is categorical, it's not possible to average incomes across different categories, however, the proportion of people with income over 50K across different categories can be compared.

Since there are many variables to explores, I focus on a few. The following questions were used as a guide for the data exploration.

Questions

1. Age

- What age group dominates the workforce? Which age group has the highest proportion of over 50K earners?
- Age can often determine how long a person has been in the workforce, and hence a higher income since as people stay longer in the workfoece, they advance in their careers, often inccreasing their income. Do people in different age groups have different income? Do older workers earn more than younger workers?

2. Sex

- The dataset is distributed between males and females workers; what proportion of each sex is represented in the data?
- Do the proportions represent the gender composition in the the general U.S. population?
- Which sex tends to have higher income?

3. Level of education

- Level of education can also be a strong determinant of one's income. Does the data suggest that people with higher levels of education have higher income?

4. Capital Gain/Capital Loss

- Can the amount of capital gain or loss indicate an individual's income?

Data Distribution

Prevalance of people earning over 50K. Out of the 32560 records in the training set, only 24% (7841) has income over 50K. Therefore, this sample has a much lower percentage of people earning over 50K and does not exactly reflect the current US income distribution which has about 45% of people earning over 50K¹.

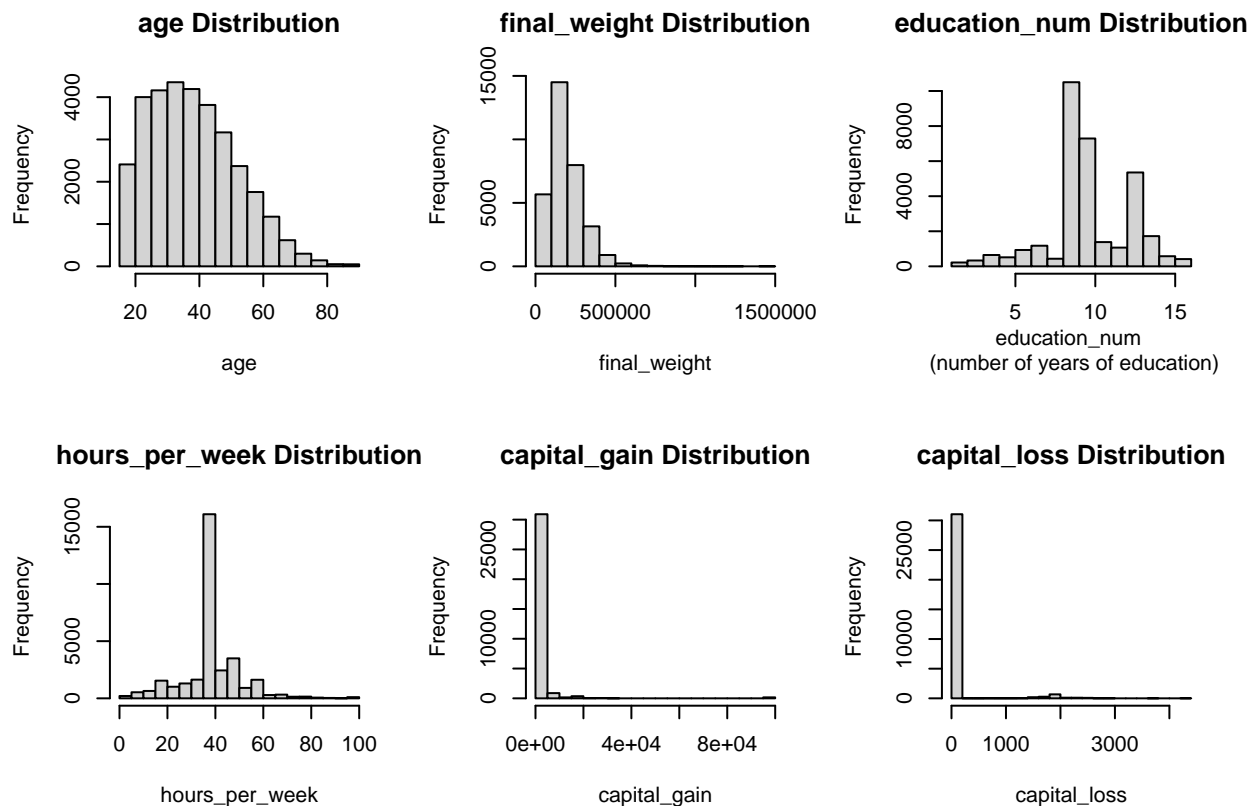
¹https://www.cs.mcgill.ca/~rwest/wikispeedia/wpcd/wp/h/Household_income_in_the_United_States.htm

Numeric Variables Distribution I start by visualizing the distribution of the numeric features.

```
adult_income %>% select(where(is.numeric)) %>% summary()
```

age	final_weight	education_num	capital_gain
Min. :17.00	Min. : 12285	Min. : 1.00	Min. : 0
1st Qu.:28.00	1st Qu.: 117832	1st Qu.: 9.00	1st Qu.: 0
Median :37.00	Median : 178363	Median :10.00	Median : 0
Mean :38.58	Mean : 189782	Mean :10.08	Mean : 1078
3rd Qu.:48.00	3rd Qu.: 237055	3rd Qu.:12.00	3rd Qu.: 0
Max. :90.00	Max. :1484705	Max. :16.00	Max. :99999

capital_loss	hours_per_week	capital_gain_loss
Min. : 0.00	Min. : 1.00	Min. : -4356.0
1st Qu.: 0.00	1st Qu.:40.00	1st Qu.: 0.0
Median : 0.00	Median :40.00	Median : 0.0
Mean : 87.31	Mean :40.44	Mean : 990.3
3rd Qu.: 0.00	3rd Qu.:45.00	3rd Qu.: 0.0
Max. :4356.00	Max. :99.00	Max. :99999.0



In general there is a wide range of values for the different variables.

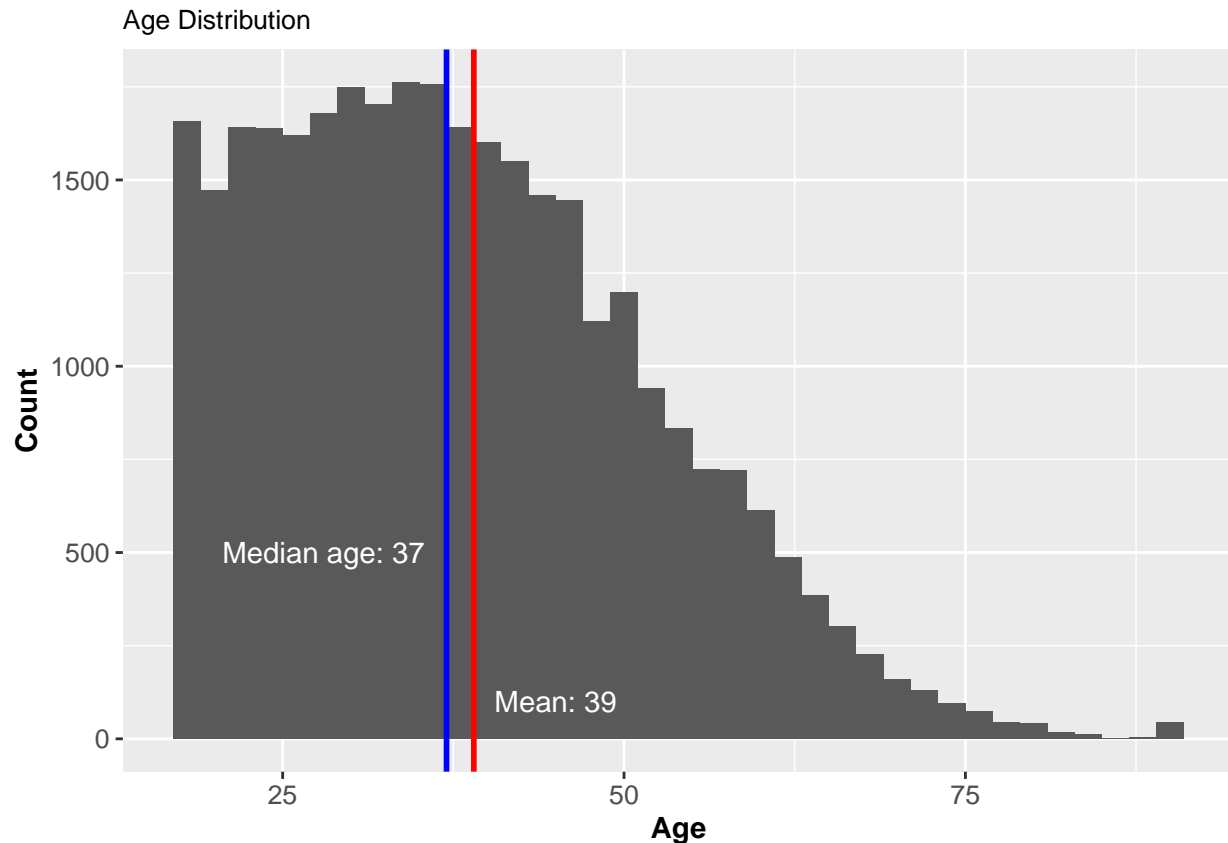
Among the six numeric variables, variable *final_weight*, has the widest range (12285 - 1484705). This is followed by the *capital_gain* and *capital_loss*; all three variables are also very highly right-skewed.

Variables *capital_gain* and *capital_loss* both have a mean and median of zero and a very small averages, reflecting on the large number of records having a value of zero (no capital gain or loss). In fact for *capital_gain*, only 8% reporting gains and 92% reported a value of zero.

Variable *education_num* is the more normally distributed than the rest, followed by *age*, and *hours_per_week*, even though *age* is still right-skewed.

The *hours_per_week* variable is a little bit of bell curve with a mean and median being same 40, meaning the majority of people are working fourty hours a week. However, one surprising thing is the presence of near 100hrs per week for some individuals. Although not a significant proportion, it is important to note that 789 individuals in this dataset work over seventy hours or more per work.

The *age* summary statistics and the histogram show a right-skewed distribution with a mean of 39 and median of 37 years. This indicates, as expected, that the bulk of the workforce is concentrated in the younger age groups².



It can be seen from the summary statistics that while the minimim age in the workforce is 17 we see some ninety year olds still in the workforce; hopefully these seniors are not drawn to work by the need for income but rather by the need to stay active.

²<https://www.bls.gov/careeroutlook/2017/article/older-workers.htm>

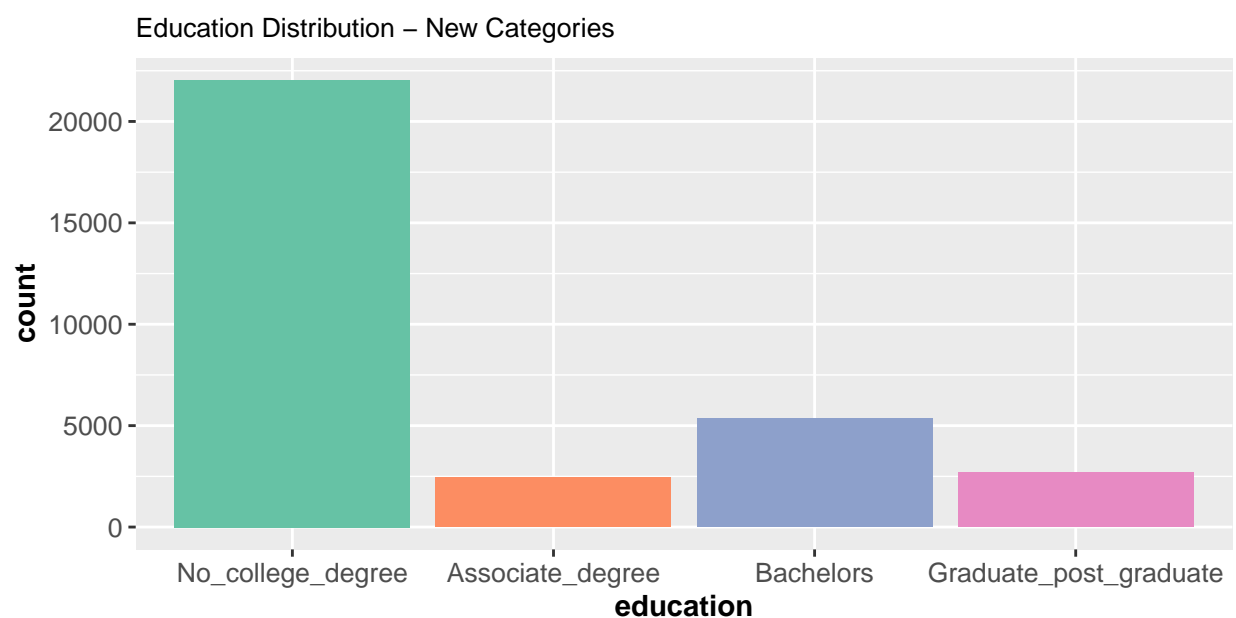
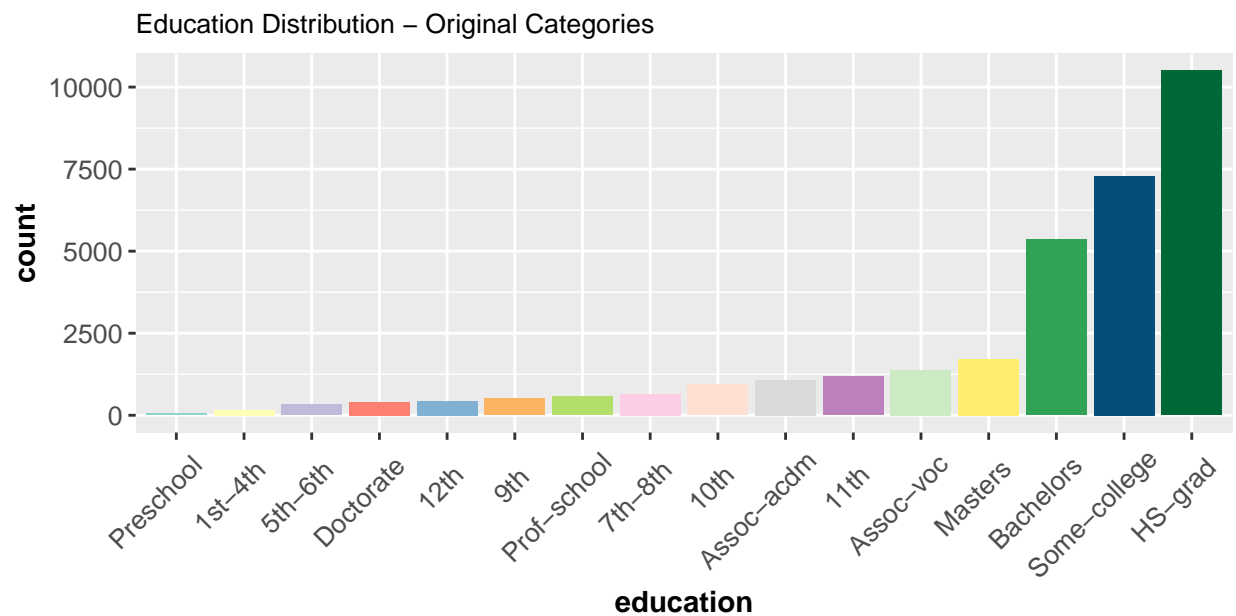
Categorical Variable Distributions Race and gender. With race, the largest proportion is White 85%. The rest is divided among Black (10%), and the others.

When it comes to gender, the sample is predominantly *male* which makes up 67% of the sample; the rest is *female*.

Education. Education is one of the categorical variables with enourmous amount of space (16 levels). The categories range from Preschool to Doctorate degree. High school diploma is the largest category (32%), followed by *Some-college* (22%), and *Masters* (16%).

In addition to *education* categorical variable, the *education_num* has the number of years of education for each individual. In essence these two vaiables may be reporting the same information.

When the categories are collapsed into more broader categories with fewer levels, it becomes even more clear that the vast majority of the individuals have no college degree. This may explain why the majority of the individuals earn less than 50K.

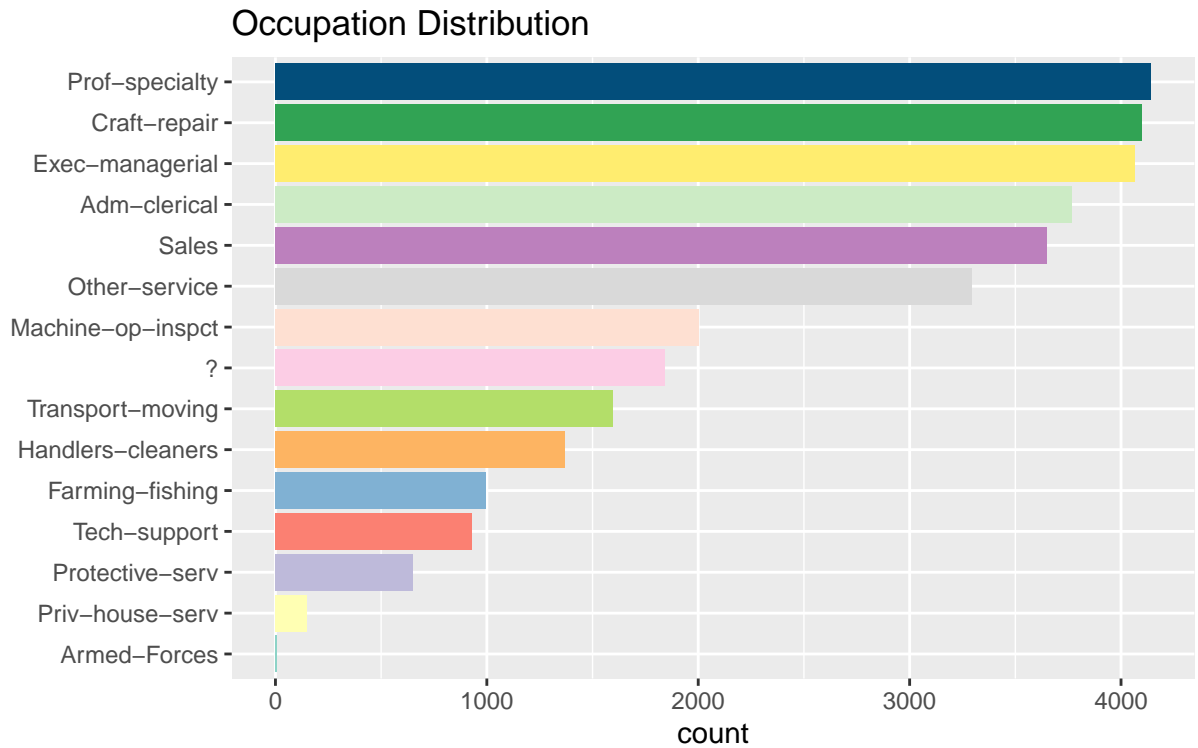


Workclass. The *workclass* variable is more like job sectors, has eight categories and one missing or not specified (?) category. The largest category is Private (70%), the rest is split among the other categories. There are some outliers (Never-worked, and Without-pay) with a counts of 7 and 14 respectively.

Occupation. Another import variable is *occupation*, a categorical variable with 15 levels, one missing (?) category.

Table 3: Occupation by Category - Orginal Categories

Category	Count
?	1843
Adm-clerical	3769
Armed-Forces	9
Craft-repair	4099
Exec-managerial	4066
Farming-fishing	994
Handlers-cleaners	1370
Machine-op-inspct	2002
Other-service	3295
Priv-house-serv	149
Prof-specialty	4140
Protective-serv	649
Sales	3650
Tech-support	928
Transport-moving	1597



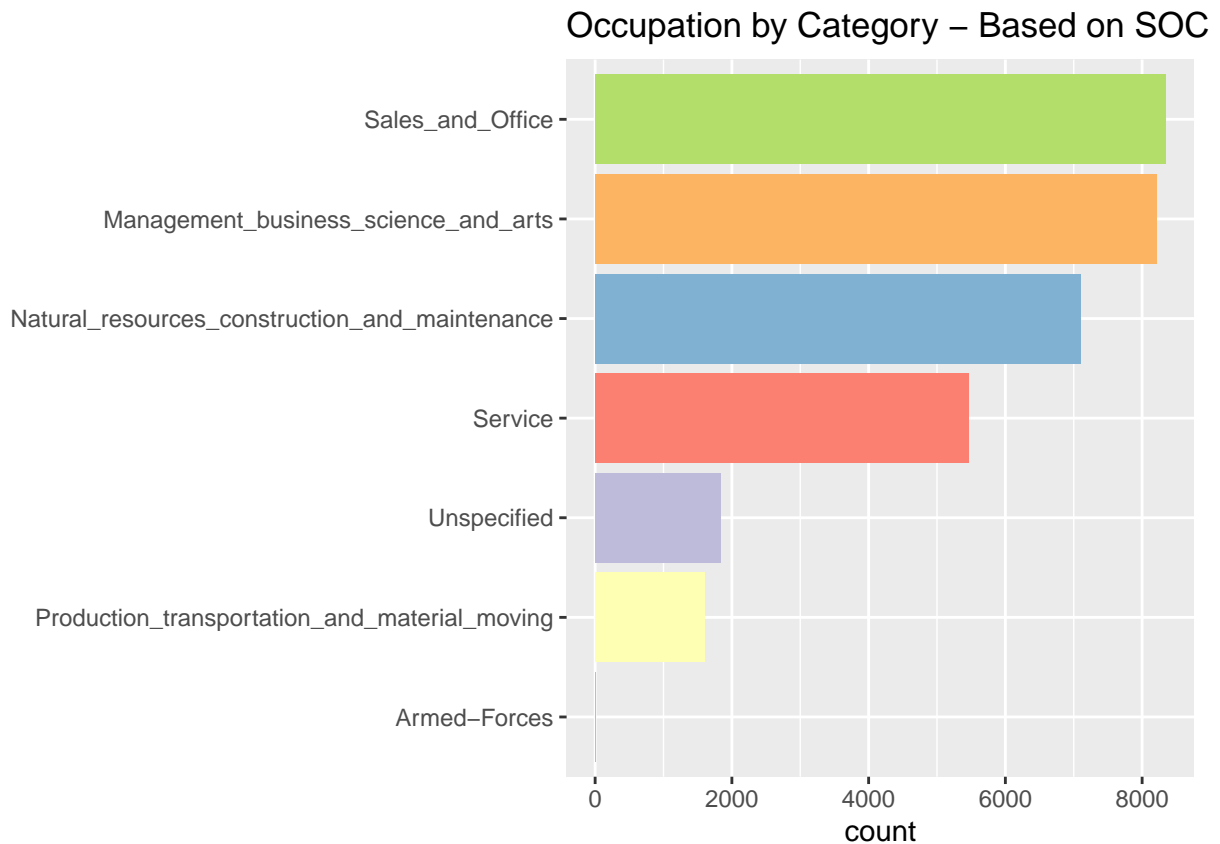
For visualization purposes, I merged the *occupation* categories into new categories based on the occupation classifications defined by US Bureau of Labor Statistics (BLS). These categories (classifications) are known as the Standard Occupation Classifications (SOC) ³.

The SOC defines many categories but the following are the ones used:

- Management_business_science_and_arts
- Service
- Sales_and_Office
- Natural_resources_construction_and_maintenance
- Production_transportation_and_material_moving

Table 4: Occupation by Category - Based on SOC Classifications

Category	Count
Unspecified	1843
Sales_and_Office	8347
Armed-Forces	9
Natural_resources_construction_and_maintenance	7095
Management_business_science_and_arts	8206
Service	5463
Production_transportation_and_material_moving	1597

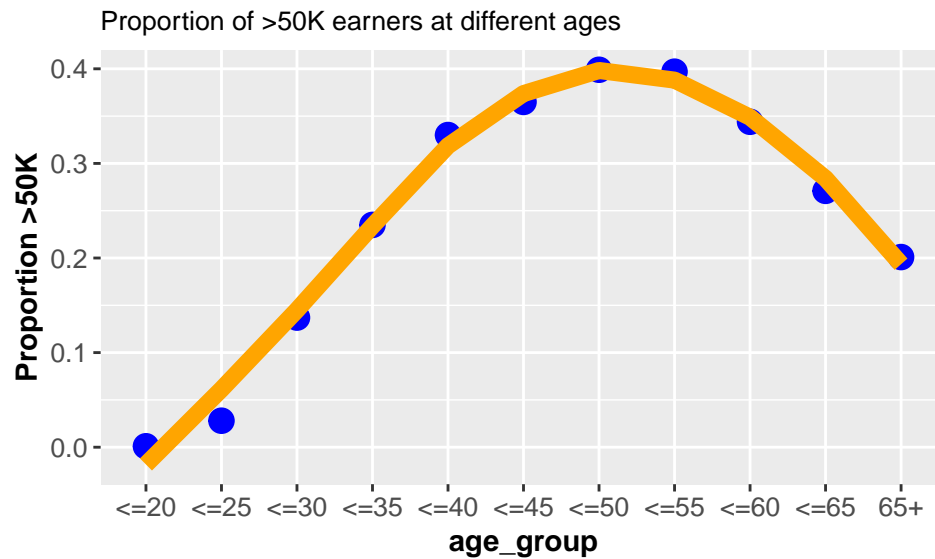


³For more information on SOC occupational classifications, see <https://www.bls.gov/soc/>

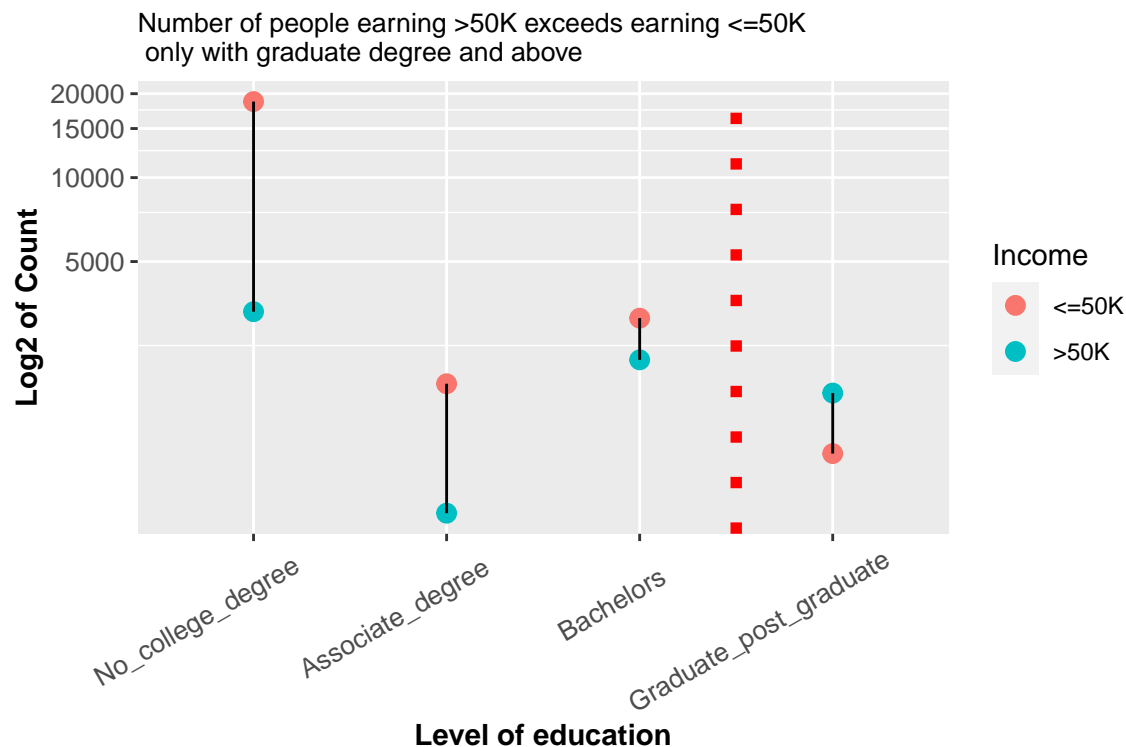
Trends

In this section I will explore the data in an attempt to find some trends.

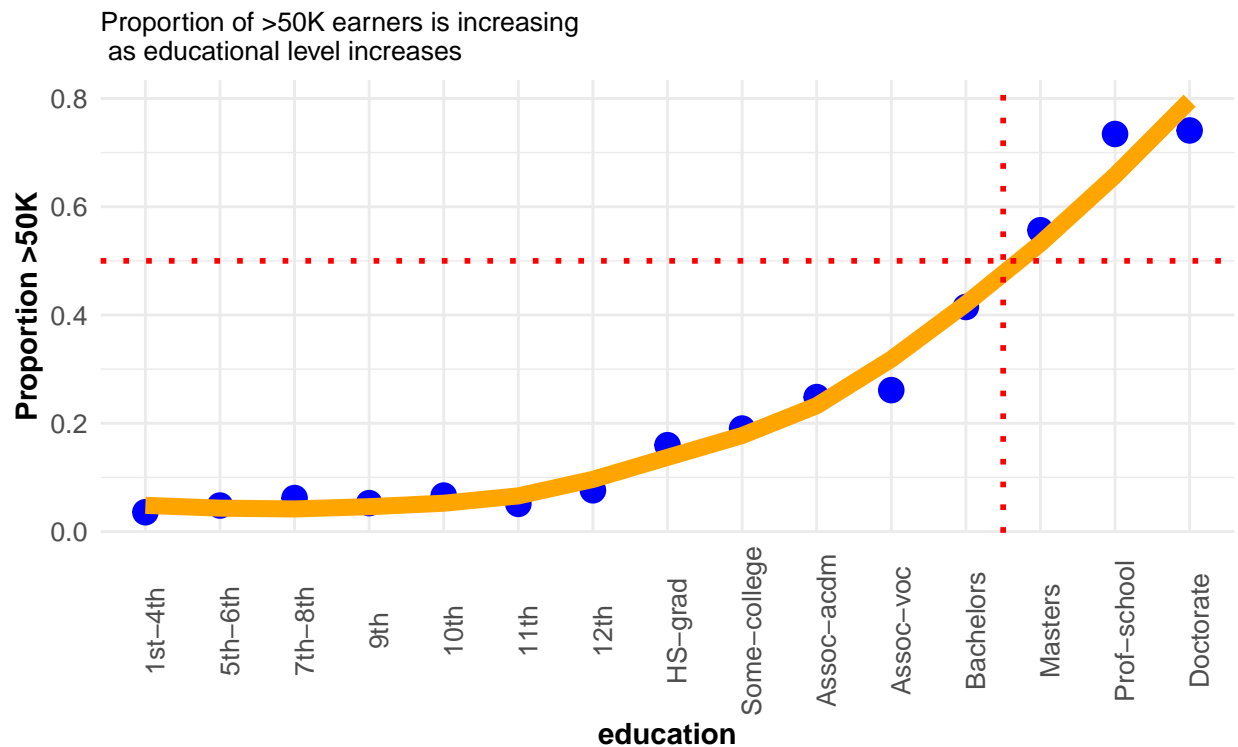
Income and Age. One of the questions asked earlier was whether age can indicate an individual's income level. The graph below shows that indeed, overall, there is a positive relationship. The proportion of individuals earning over 50K increases with age and peaks around the fifties, before starting to decline.



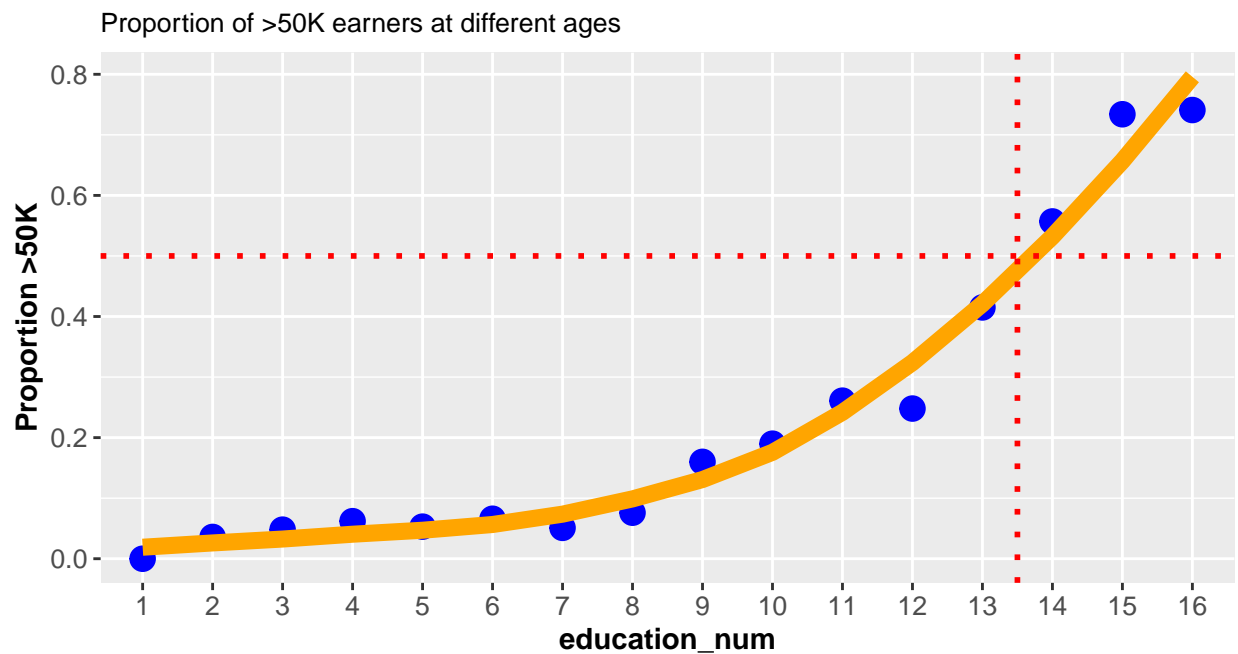
Income and education. It is generally accepted that income level is associated with educational attainment; the higher the educational level, the higher the income. Does the data support this belief? Based on the data, at every educational levels (except graduate and post graduate), the number of individuals earning >50K is lower than those earning <=50K.



However, as the graph below shows, the proportion of individuals earning >50K is actually increasing and the rate gets steeper after an individual attains some college level education. Therefore, education seems to have a positive effect on income.



It is important to note that the graph above can be generated using *education_num* instead of the categorical variable *education* (see chart below). This is because both *education_num* and *education* have the same information. Since these two variables may be correlated, *education_num* will be excluded in final analysis.



Income by gender . Gender and income inequality has always been a topic of concern in almost every society. This section shows a simple analysis of how income differs by gender in the dataset. The table below shows that in general, a higher proportion of males earn higher income than females.

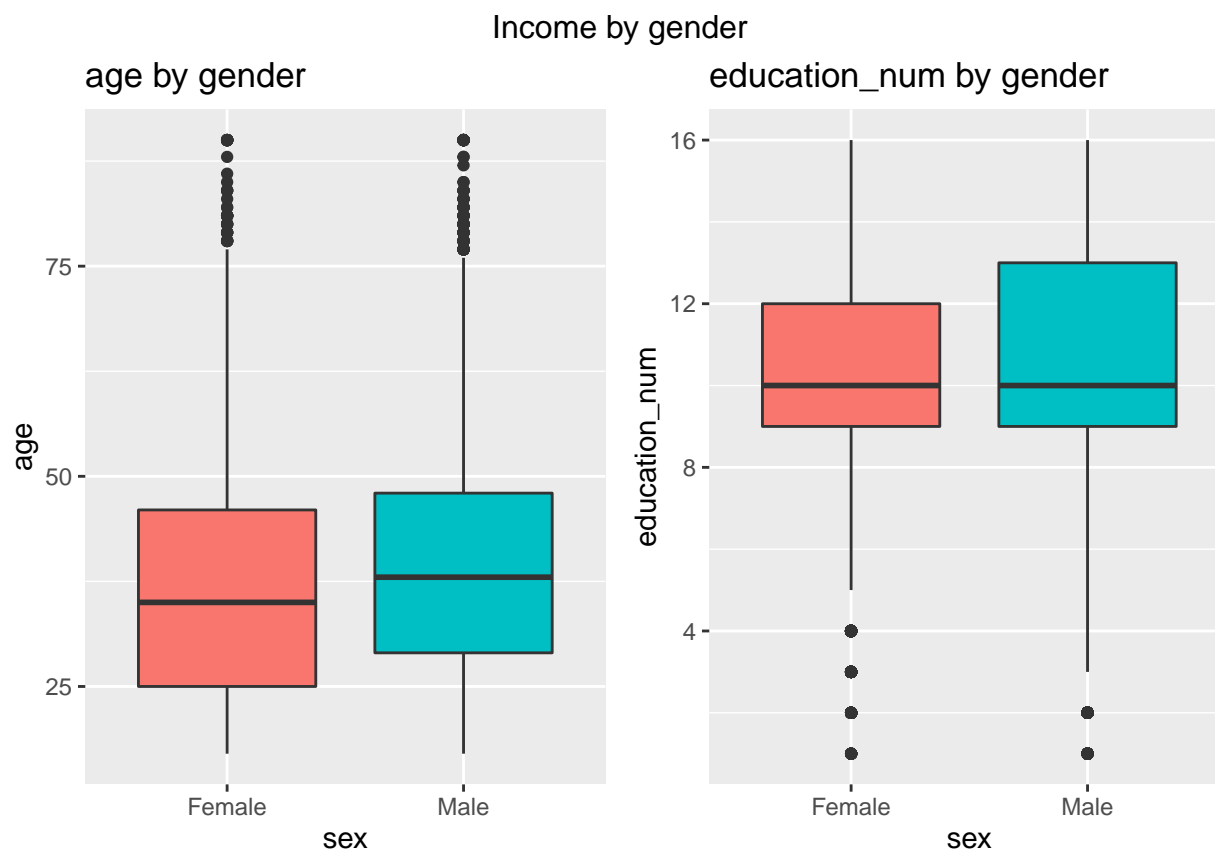
Table 5: Income and Gender

Gender	Total	Total over50K	Proportion >50K
Female	10771	1179	0.11
Male	21789	6662	0.31

Education wise, the females are equally educated as the men, but the average age for females is a bit lower than that of the males, meaning there are more younger women in the workforce (probably earning lower salaries), but am not sure if that explains the difference.

Table 6: Gender, Avg Age, and Education

Gender	Avg Age	Avg Education Yrs
Female	37	10
Male	39	10



Income and capital gain/capital loss. Another set of variables I wanted to explore is the *capital_gain* and *capital_loss*. In general a vast majority of individuals did not report a gain or loss. And even with those that reported a gain, there is a wide range for the values reported. Do people who reported capital gain have higher income or vice versa?

The following chart and tables explore if capital gain or loss can provide an indication of one's income level.

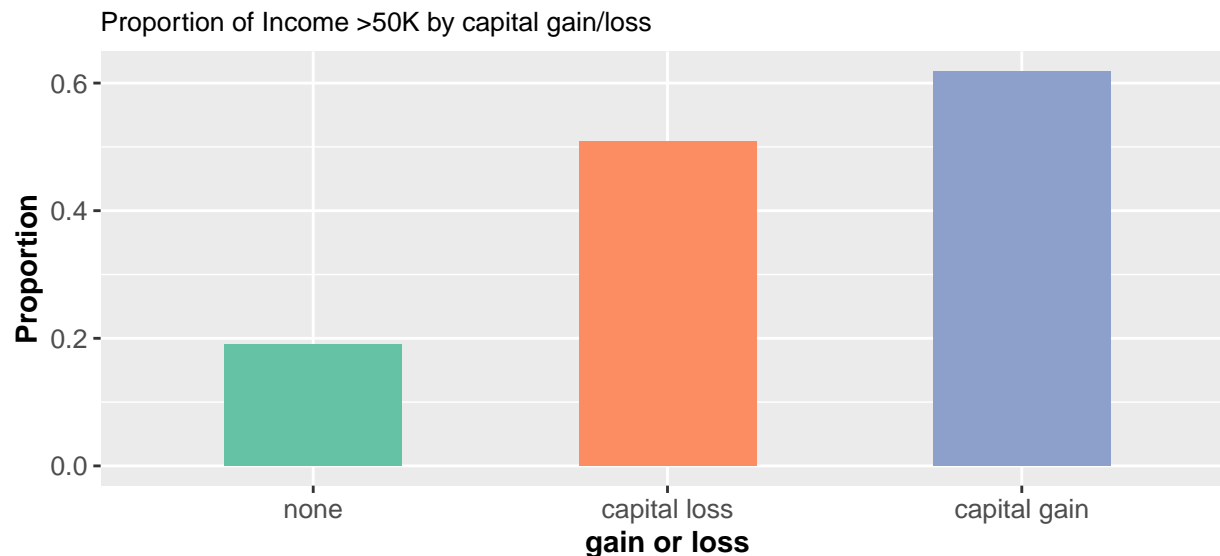


Table 7: Proportion of Income over50K by Capital Gain

capital_gain_level	total	over50K	proportion
none	29849	6164	0.2065061
<=10K	1941	921	0.4744977
<=20K	517	511	0.9883946
>20K	253	245	0.9683794

Table 8: Proportion of Income over50K by Capital Loss

capital_loss_level	total	over50K	proportion
none	31041	7068	0.2276988
<=1K	36	2	0.0555556
<=2K	1158	629	0.5431779
<=3K	314	139	0.4426752
>3K	11	3	0.2727273

It can be noted from the chart above that individuals who reported no capital gain or loss have a higher probability of earning <=50K. Individuals reporting capital gains have a higher probability of earning >50K. From the tables, it can be seen that Individuals with a gain of 20K and above are almost guaranteed to earn >50K. In fact, the maximum capital gain reported, 99999, is reported by 159 individuals; out of those individuals, not a single one reported an income <=50K.

```
adult_income %>% filter(capital_gain==max(capital_gain) & Income=="<=50K" ) %>% nrow()
```

[1] 0

Analysis

Approach

This section presents the analysis, approach and results. The proposed approach is outlined below:

0. Data partitioning (train/test)
1. Data Engineering, Transformation & Preprocessing
2. Training, Optimization & Model Selection
3. Final model and final prediction on validation set

Data Partitioning (train/test)

The training data (adult_income_train dataset) is split into train/test (90/10 partitions).

Data Engineering, Transformation & Preprocessing

The datasets used in this project had been cleaned by the original contributors, however, a few tasks were performed to prepare the data for the final analysis.

- 1. replace question marks in categorical data.** The question mark “?” values were replaced with the value *Unspecified* in the three variables that had this value.
- 2. remove special characters.** Remove special characters like the ampersand (&) and hyphen (-) from categorical values.
- 3. remove period from the values in the response variable Income.** The validation dataset had period “.” in the values of the response variable, Income.

```
unique(adult_income_validation$Income)
```

```
[1] "<=50K." ">50K."
```

- 4. Convert the values in the response variable to Y,N.** Before the analysis, the values in the response (Income) was converted to Y/N: (Y equals >50K, N: equals <=50K) and with Y being set as the positive.
- 5. convert all categorical variables into dummy variables.** All categorical variables were converted into numeric (0,1) dummy variables using one-hot encoding from the caret package.
- 6. log transformations.** Some variables (*age* and *final_weight*) were log2 transformed before scaling.
- 7. new column for capital gain/loss.** The *capital_gain* and *capital_loss* variables were combined into a single column called *capital_gain_loss*. The values then range from positive (capital gain) to negative (capital loss). In addition, values were all adjusted to remove zeros and then log2 transformed.
- 6. scale all numeric variables.** After all the transformation, the variables were numeric; the final step was to scale them before the analysis.

The tasks for performing the transformation are all placed in R functions so they can be reused.

```
#####
#
# PREPROCESSING PIPELINE
# TAKE ALL DATASETS THROUGH THE SAME TRANSFORMATION
#
#####
# 1. rename some categorical values
training_set_final <- fnManualTransformations(train_set)
test_set_final     <- fnManualTransformations(test_set)

# 2. create caret preprocessor for dummy variables
caret_dummy_vars_preprocessor <- fnCreateCaretDummyVarsPreproc(training_set_final)
training_set_final <- fnApplyCaretDummyVarsPreproc(training_set_final,
                                                    caret_dummy_vars_preprocessor)

test_set_final     <- fnApplyCaretDummyVarsPreproc(test_set_final,
                                                    caret_dummy_vars_preprocessor)

# 3. cosmetic stuff
# convert response var to Y,N, and set the positive class to Y (over50K)
training_set_final <- fnEncodeResponseVar(training_set_final)
test_set_final     <- fnEncodeResponseVar(test_set_final)

# 4. perform final caret preprocess - center and scale
caret_scale_preprocessor <- fnCreateCaretScalePreproc(training_set_final)
training_set_final <- fnApplyCaretScalePreproc(training_set_final,
                                                caret_scale_preprocessor)

test_set_final     <- fnApplyCaretScalePreproc(test_set_final,
                                                caret_scale_preprocessor)
```

After all the transformations are performed on the datasets, the training data looks are below:

```
> glimpse(training_set_final)
Rows: 29,303
Columns: 92
$ over50K          <fct> N, N, N, N, N, N, Y, Y, Y, Y, Y, N, N, N, N, N, Y, Y, N, N, N, N, Y, N, N,
$ age             <dbl> 0.89547830, 0.13429748, 1.05709354, -0.71271217, 0.06033015, 0.83944387, 1.004
$ workclass.Federal_gov <dbl> -0.1747132, -0.1747132, -0.1747132, -0.1747132, -0.1747132, -0.1747132, -0.174
$ workclass.Local_gov  <dbl> -0.2627185, -0.2627185, -0.2627185, -0.2627185, -0.2627185, -0.2627185, -0.262
$ workclass.Never_worked <dbl> -0.01545744, -0.01545744, -0.01545744, -0.01545744, -0.01545744, -0.01545744, -0.015
$ workclass.Private     <dbl> -1.517615, 0.658906, 0.658906, 0.658906, 0.658906, 0.658906, -1.517615, 0.658
$ workclass.Self_emp_inc <dbl> -0.1878674, -0.1878674, -0.1878674, -0.1878674, -0.1878674, -0.1878674, -0.187
$ workclass.Self_emp_not_inc <dbl> 3.4410064, -0.2906027, -0.2906027, -0.2906027, -0.2906027, -0.2906027, 3.44100
$ workclass.State_gov   <dbl> -0.2037467, -0.2037467, -0.2037467, -0.2037467, -0.2037467, -0.2037467, -0.203
$ workclass.Unspecified <dbl> -0.2437167, -0.2437167, -0.2437167, -0.2437167, -0.2437167, -0.2437167, -0.243
$ workclass.Without_pay <dbl> -0.02024026, -0.02024026, -0.02024026, -0.02024026, -0.02024026, -0.02024026, -0.020
$ final_weight        <dbl> -1.036941335, 0.473299420, 0.607893900, 1.188860965, 0.913772759, 0.001205001,
$ education_num       <dbl> 1.1357831, -0.4215358, -1.2001953, 1.1357831, 1.5251128, -1.9788547, -0.421535
$ marital_status.Divorced <dbl> -0.3978788, 2.5132427, -0.3978788, -0.3978788, -0.3978788, -0.3978788, -0.3978
$ marital_status.Married_AF_spouse <dbl> -0.02677946, -0.02677946, -0.02677946, -0.02677946, -0.02677946, -0.02677946, -0.02677946,
$ marital_status.Married_civ_spouse <dbl> 1.081995, -0.924187, 1.081995, 1.081995, 1.081995, -0.924187, 1.081995, -0.924
```

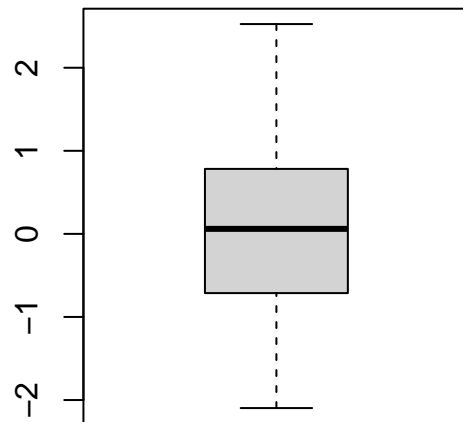
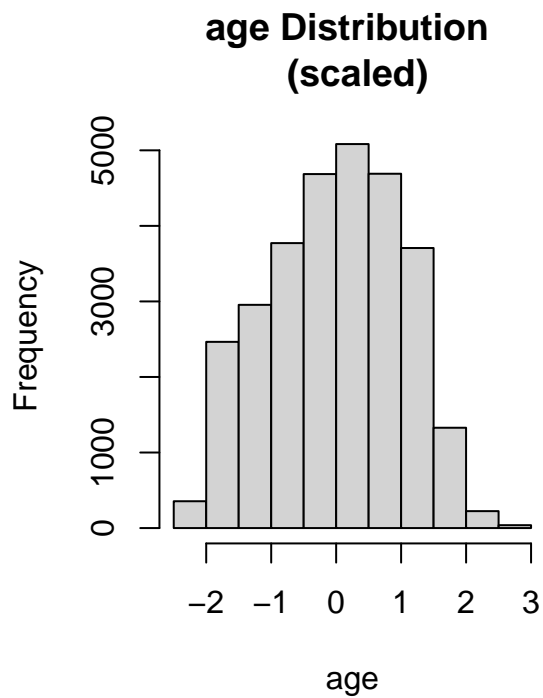
Figure 2: training data after transformation and scaling

The summary, and scatterplots, and bpxplots below show the distribution of the new scaled variables.

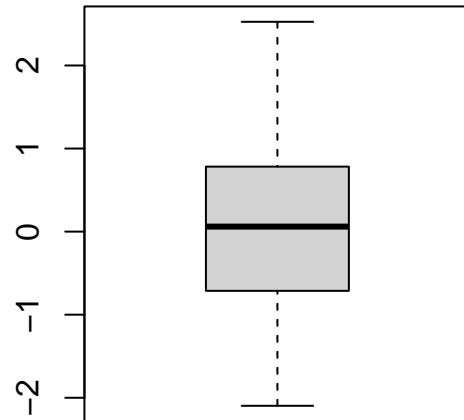
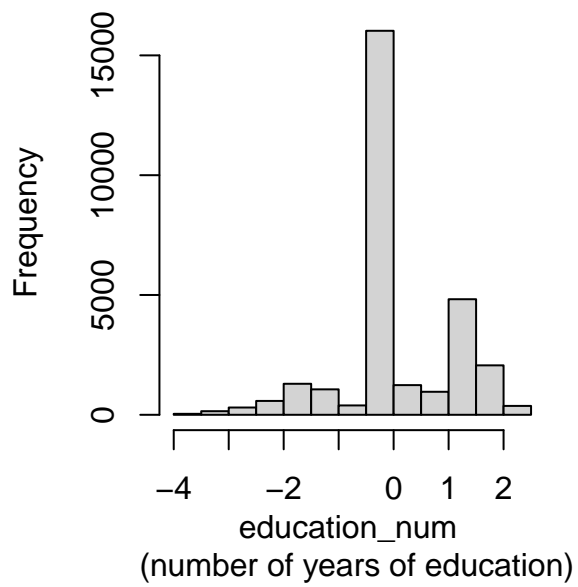
```
training_set_final %>%
  select(c(age, final_weight, education_num, hours_per_week, capital_gain_loss)) %>%
  summary()
```

age	final_weight	education_num	hours_per_week
Min. :-2.09672	Min. :-4.0766	Min. :-3.53617	Min. :-3.19091
1st Qu.: -0.71271	1st Qu.: -0.4860	1st Qu.: -0.42154	1st Qu.: -0.03469
Median : 0.06033	Median : 0.1700	Median : -0.03221	Median : -0.03469
Mean : 0.00000	Mean : 0.0000	Mean : 0.00000	Mean : 0.00000
3rd Qu.: 0.78225	3rd Qu.: 0.6223	3rd Qu.: 0.74645	3rd Qu.: 0.36996
Max. : 2.52577	Max. : 3.5054	Max. : 2.30377	Max. : 4.74011

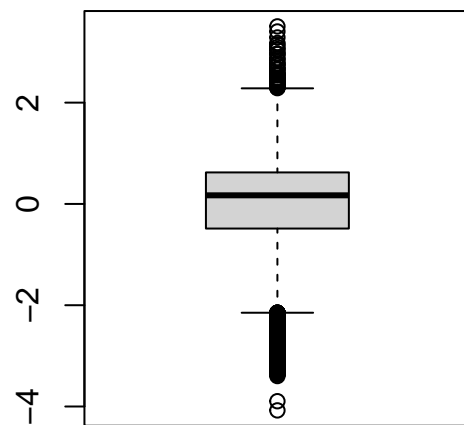
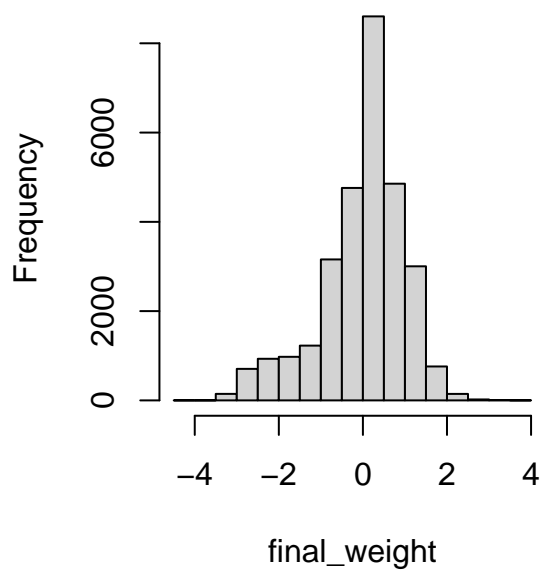
capital_gain_loss
Min. :-22.1528
1st Qu.: -0.1527
Median : -0.1527
Mean : 0.0000
3rd Qu.: -0.1527
Max. : 8.1858



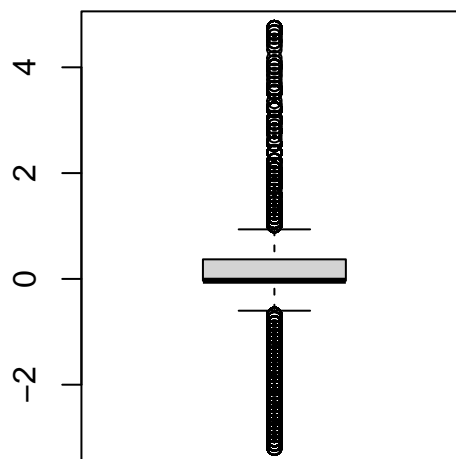
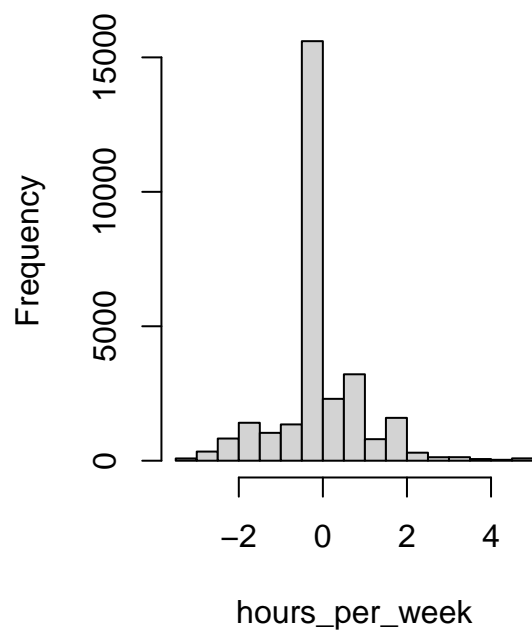
**education_num Distribution
(scaled)**



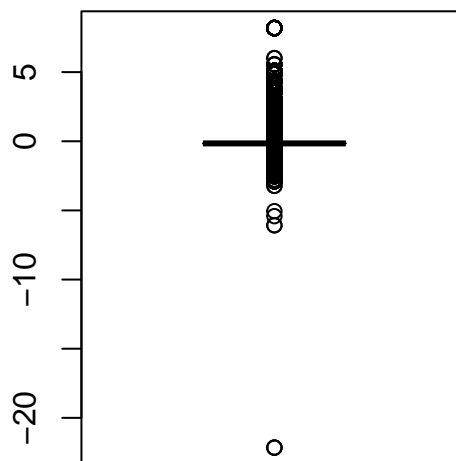
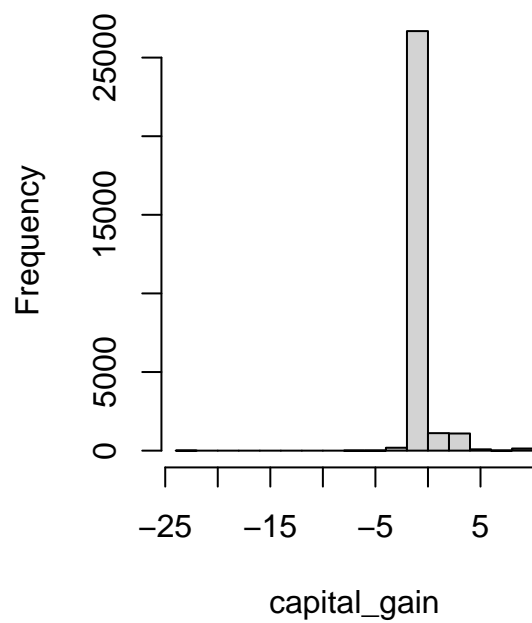
**final_weight Distribution
(scaled)**



**hours_per_week Distribution
(scaled)**



**capital_gain/loss Distribution
(scaled)**



Training, Optimization, & Final Model and Results

Datasets The following table lists the datasets as used in the final part of the project.

Table 9: Final datasets for analysis and validation

data	numrows
training_set_final	29303
test_set_final	3257
validation	16280

Cross Validation and Sampling. To avoid overfitting, a 5-fold cross validation was used while training the models using a sample size of 5000 from the training set.

The following parameters are used to define a common training control for all models:

```
cv_control <- trainControl(method = "cv",
                           number = 5,
                           summaryFunction=twoClassSummary,
                           classProbs=TRUE,
                           savePredictions = T)
```

The *summaryFunction* is set to *twoClassSummary*, *classProbs=TRUE*, both are required when using ROC as the performance metric. *savePrediction* is set to *TRUE* so caret will retain the predictions from the final model, which can be used later.

Performance metric. In this projet, I use *AUC* (Area Under the Curve) as the metric for selecting best models during cross validation. In the confusion matrix, use Kappa instead of accuracy when selecting best model. Kappa is often preferred to accuracy when dealing with imbalanced dataset which is the case in this project. In th caret training function, the *metric* will be set to **ROC**, to indicate to caret to select best model using *AUC* rather than Accuracy.

KNN Models

First, I trained two KNN models and picked the best performing one. My approach was, first, I trained a model using caret defaults (knn_model_1). In knn, there is only one parameter to tune k. In the default model, I used a tuneLength of 10 to allow caret to select 10 values for k. This model provided a *best* value for k from the cross validation. I fit that model to the entire training set, use the fit to make predictions on the test.

I take the *best* k from the first model, and expanded around it to provide more tuning values for k and used it to train another model (knn_model_2). I fitted the new model on the entire training set, and made another predictions on test. I then compare the two models.

```
knn_model_1 <- train(
  over50K ~ .,
  method = "knn",
  metric="ROC",
  data = train_sample,
  trControl = cv_control,
  tuneLength=10
)
```

```
knn_k_tuning_grid <- data.frame(k = c(seq(1, 39, 2), 50, 100, 150, 200, 250, 350))
knn_model_2 <- train(
  over50K ~ .,
  method = "knn",
  metric="ROC",
  data = train_sample,
  trControl = cv_control,
  tuneGrid = knn_k_tuning_grid
)
```

The following shows the two models, the cross validation results, and confusion matrix.

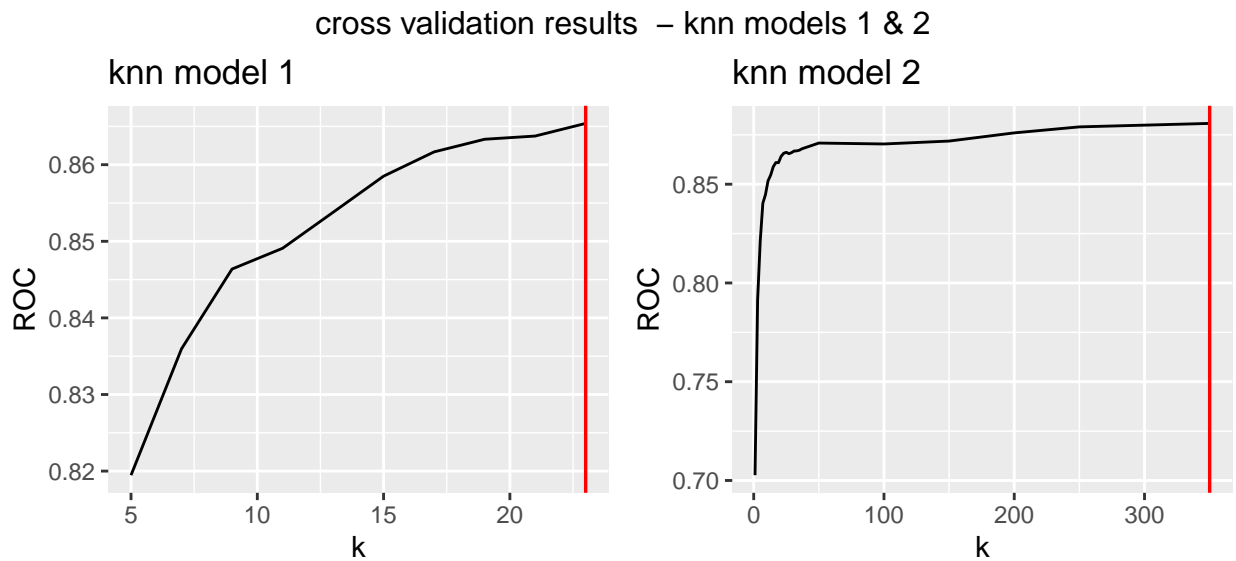
Table 10: knn model 1 - cross validation results

k	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
5	0.8194645	0.5134454	0.9020997	0.0186233	0.0531807	0.0079068
7	0.8359619	0.5033613	0.9107612	0.0162358	0.0493049	0.0093720
9	0.8463806	0.4941176	0.9154856	0.0176524	0.0567305	0.0080149
11	0.8490863	0.4924370	0.9204724	0.0159806	0.0580228	0.0086355
13	0.8537848	0.4907563	0.9236220	0.0162496	0.0700812	0.0063210
15	0.8585048	0.4857143	0.9236220	0.0140649	0.0549120	0.0129383
17	0.8616726	0.5000000	0.9223097	0.0145772	0.0601589	0.0106372
19	0.8633241	0.5067227	0.9257218	0.0137528	0.0584925	0.0103998
21	0.8637398	0.5025210	0.9272966	0.0151050	0.0527809	0.0114256
23	0.8653924	0.4966387	0.9267717	0.0150496	0.0546542	0.0099253

Table 11: knn model 2 - cross validation results

k	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
1	0.7026633	0.5470588	0.8582677	0.0114777	0.0069041	0.0223290
3	0.7914014	0.5336134	0.8916010	0.0090559	0.0194824	0.0125395
5	0.8211099	0.5260504	0.9097113	0.0127138	0.0227046	0.0170957
7	0.8403725	0.5193277	0.9146982	0.0161662	0.0274881	0.0208740
9	0.8447054	0.4966387	0.9167979	0.0133604	0.0256268	0.0184802
11	0.8517198	0.4924370	0.9196850	0.0130641	0.0306464	0.0182410
13	0.8546130	0.4915966	0.9204724	0.0106164	0.0289581	0.0186887
15	0.8588456	0.4983193	0.9175853	0.0117789	0.0253149	0.0189586
17	0.8609806	0.5025210	0.9238845	0.0108863	0.0312454	0.0171108
19	0.8608692	0.5075630	0.9225722	0.0110679	0.0407801	0.0189268
21	0.8640817	0.5109244	0.9207349	0.0116675	0.0339531	0.0181036
23	0.8656604	0.5025210	0.9225722	0.0119711	0.0406934	0.0218613
25	0.8661616	0.5126050	0.9228346	0.0122705	0.0460271	0.0226050
27	0.8654503	0.5100840	0.9225722	0.0120760	0.0435234	0.0238217
29	0.8659791	0.5067227	0.9225722	0.0105097	0.0488733	0.0219007
31	0.8667979	0.4991597	0.9223097	0.0116144	0.0434016	0.0218298
33	0.8668850	0.5084034	0.9225722	0.0114330	0.0489996	0.0204783
35	0.8672010	0.5058824	0.9228346	0.0114747	0.0458927	0.0210465
37	0.8679277	0.5008403	0.9241470	0.0104992	0.0480353	0.0228701
39	0.8683771	0.4957983	0.9249344	0.0107148	0.0431568	0.0230015
50	0.8708358	0.4932773	0.9304462	0.0126514	0.0401475	0.0194208
100	0.8703776	0.4932773	0.9291339	0.0133842	0.0515969	0.0204783
150	0.8718520	0.4966387	0.9301837	0.0122222	0.0307900	0.0219674
200	0.8759964	0.4865546	0.9367454	0.0130689	0.0441076	0.0203391
250	0.8790307	0.4798319	0.9393701	0.0136318	0.0403449	0.0196281
350	0.8807974	0.4680672	0.9451444	0.0140097	0.0320817	0.0195181

The following show the graphs of the cross validation results.



From the cross validation results in first model, it appear k is still increasing and might even yield a better result if we increase k. But from the second model, it's clear that increasing k does not yield better results. In fact the first model using caret default is better. The following details from the confusion matrix show how both models perform on the test set.

ConfusionMatrix on test set – knn models 1 & 2

	term	estimate		term	estimate
1	accuracy	0.8397298	1	accuracy	0.8357384
2	kappa	0.5234003	2	kappa	0.4960396
3	sensitivity	0.5490446	3	sensitivity	0.4993631
4	specificity	0.9320388	4	specificity	0.9425566
5	pos_pred_value	0.7195326	5	pos_pred_value	0.7340824
6	neg_pred_value	0.8668172	6	neg_pred_value	0.8556739
7	precision	0.7195326	7	precision	0.7340824
8	recall	0.5490446	8	recall	0.4993631
9	f1	0.6228324	9	f1	0.5943897
10	prevalence	0.2410193	10	prevalence	0.2410193
11	detection_rate	0.1323304	11	detection_rate	0.1203562
12	detection_prevalence	0.1839116	12	detection_prevalence	0.1639546
13	balanced_accuracy	0.7405417	13	balanced_accuracy	0.7209598
14	accuracy.conf.low	0.8266734	14	accuracy.conf.low	0.8225602
15	accuracy.conf.high	0.8521730	15	accuracy.conf.high	0.8483107
16	no_information_rate	0.7589807	16	no_information_rate	0.7589807

We can see from the details above that in model 1 (left table), accuracy, kappa, and sensitivity are higher than those in model 2 (right table). Therefore, I selected knn_model_1 as best model from knn. Next I will train other models using randomForest and select the best model again from there. I will then pick from two final models, the best performing model and use as my final model for validation.

RandomForst Models

RandomForest Model 1. Using randomForest defaults for training. In this section, I used the same approach as used when training the knn models. I used a sample (5K observations) to train the models. There are two main parameters of interest to me; *mtry* (number of random features to use in each split), and *nodesize* (minimum number of nodes in the terminal nodes or leaves). Number of tree (*ntree*) can also be set for randomForest, but in this case, I let caret choose values for *ntree*.

First, I train a model using caret defaults (*rf_model_1*), then take the best *mtry* from that model, expand on it and use it to train another model (*rf_model_2*) to see if there is an *mtry* that yields a better performance. I then use the new *mtry* value and train another model (*rf_model_3*), this time searching for an optimized *nodesize*. At the end, I select the best model among these three, and compare that to the one from knn.

```
# use caret defaults to train rf model using cross validation
# use ROC as a metric to for selecting best model
fnTrainRandomForestWithDefaults <- function(train_sample){
  train(
    over50K ~ .,
    method = "rf",
    metric="ROC",
    data = train_sample,
    trControl = cv_control,
    tuneLength=10,
    allowParallel=TRUE
  )
}
```

Again I set *tuneLength* to 10 to allow caret to generate 10 random values for *mtry* (by default caret will pick three values for *mtry*).

The training model was then used to fit a model on the entire training set, and used it to make predictions on the test set. The results below show the results from the cross validation. The cross validation results show that the ROC ranges from 88% to 90%. The best cross validation model achieves ROC of 90% with *mtry* of 21, at 61% sensitivity, and 93% specificity.

Table 12: Cross validation results - model 1

<i>mtry</i>	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
2	0.8827648	0.0310924	1.0000000	0.0131190	0.0117347	0.0000000
11	0.9027691	0.5974790	0.9406824	0.0089822	0.0303570	0.0113044
21	0.9038819	0.6142857	0.9330709	0.0091404	0.0313582	0.0073655
31	0.9014414	0.6134454	0.9296588	0.0078120	0.0275523	0.0081745
41	0.9010637	0.6218487	0.9317585	0.0076637	0.0243190	0.0089489
51	0.9001312	0.6235294	0.9283465	0.0079093	0.0223520	0.0065354
61	0.8979405	0.6243697	0.9288714	0.0079417	0.0253149	0.0058248
71	0.8969849	0.6252101	0.9270341	0.0072142	0.0312171	0.0079610
81	0.8951041	0.6252101	0.9262467	0.0080539	0.0166484	0.0063888
91	0.8940134	0.6176471	0.9267717	0.0088839	0.0224308	0.0069690

```
$positive
[1] "Y"
```

```

$table
      Reference
Prediction  Y    N
      Y  462 169
      N  323 2303

$overall
      Accuracy      Kappa  AccuracyLower  AccuracyUpper  AccuracyNull
      8.489407e-01  5.574877e-01  8.361761e-01  8.610753e-01  7.589807e-01
AccuracyPValue  McNemarPValue
      9.256243e-37  5.282417e-12

$byClass
      Sensitivity      Specificity      Pos Pred Value
      0.5885350      0.9316343      0.7321712
      Neg Pred Value      Precision      Recall
      0.8769992      0.7321712      0.5885350
      F1      Prevalence      Detection Rate
      0.6525424      0.2410193      0.1418483
Detection Prevalence  Balanced Accuracy
      0.1937366      0.7600847

$mode
[1] "sens_spec"

$dots
list()

attr(,"class")
[1] "confusionMatrix"

```

From the confusion matrix, we see that the accuracy is 0.849. Sensitivity (the True Positive prediction rate) is 0.589, much lower than the Specificity (True Negative prediction rate) of 0.932. This is due to the low prevalence of the positive class, as already mentioned. Our main performance statistic, Kappa is 0.557.

Overall, without any tuning, the model performance is pretty good; Sensitivity above 59% indicates that the model is doing fairly good in predicting the positive class. The Kappa value also indicates that the model is doing fairly good in distinguishing between the positive and negative classes. The next models attempt to improve upon the first model by tuning some of the parameters.

RandomForest Model 2. Attempt to tune mtry paramaters. With this model, the same sample (5K observations from the training set) with the same cross validation were used in an attempt to tune mtry. From model 1, i the best mtry was 21 . The idea was to expand the around the mtry obtained from model 1 and see if there is a better mtry. With this approach, the best cross validation ROC for model 2 is achieved with mtry = 15.

```

# model 2 - train an rf model with cross validation and with different values
# for best mtry
fnTrainRandomforest_Tune_mtry <- function(train_sample){
  train(
    over50K ~ .,
    method = "rf",
    data = train_sample,

```

```

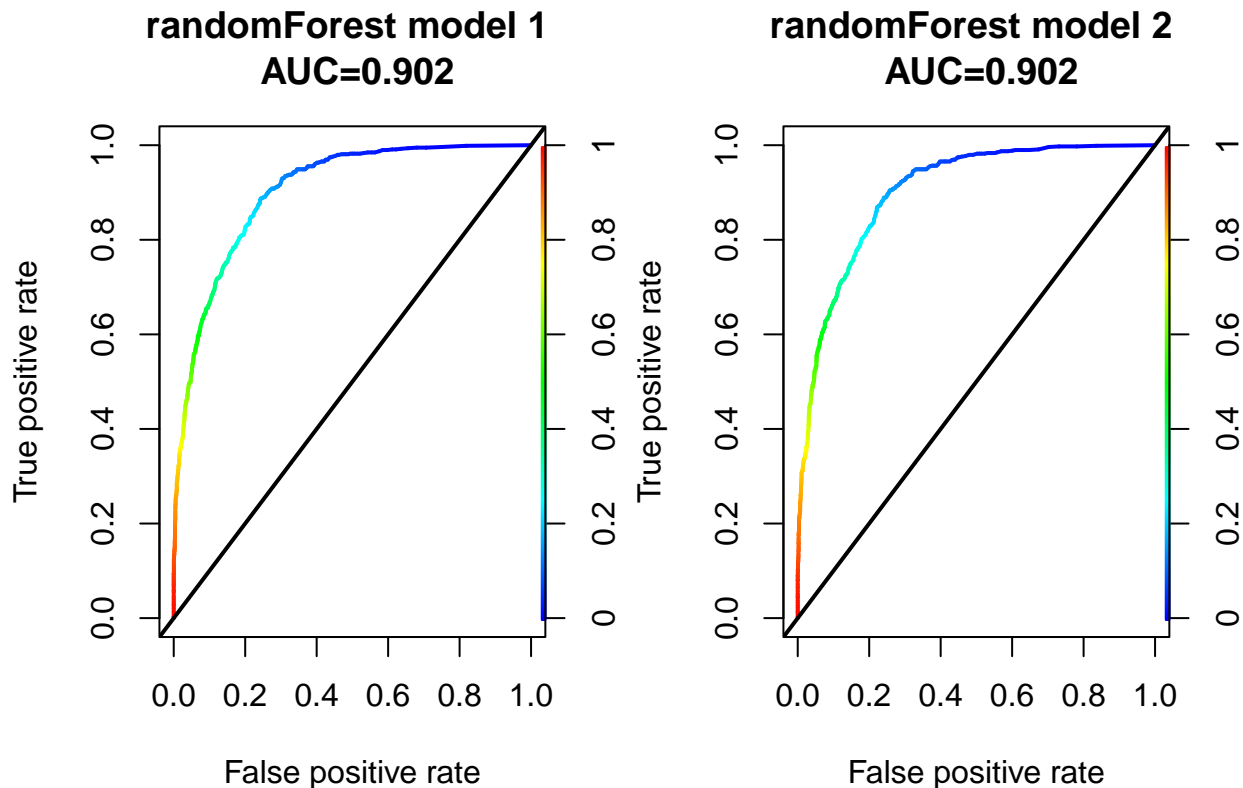
metric="ROC",
trControl = cv_control,
tuneGrid = data.frame(mtry = c(5, 10, 15, 20, 30)),
allowParallel=TRUE
)
}

```

Table 13: Cross validation results - model 2 finding best mtry

mtry	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
5	0.8973114	0.5067227	0.9545932	0.0111032	0.0260029	0.0104411
10	0.9031932	0.6058824	0.9398950	0.0085270	0.0272948	0.0114181
15	0.9060720	0.6243697	0.9391076	0.0082855	0.0394824	0.0090732
20	0.9057434	0.6285714	0.9359580	0.0083537	0.0295017	0.0111895
30	0.9039376	0.6310924	0.9296588	0.0076313	0.0188842	0.0127776

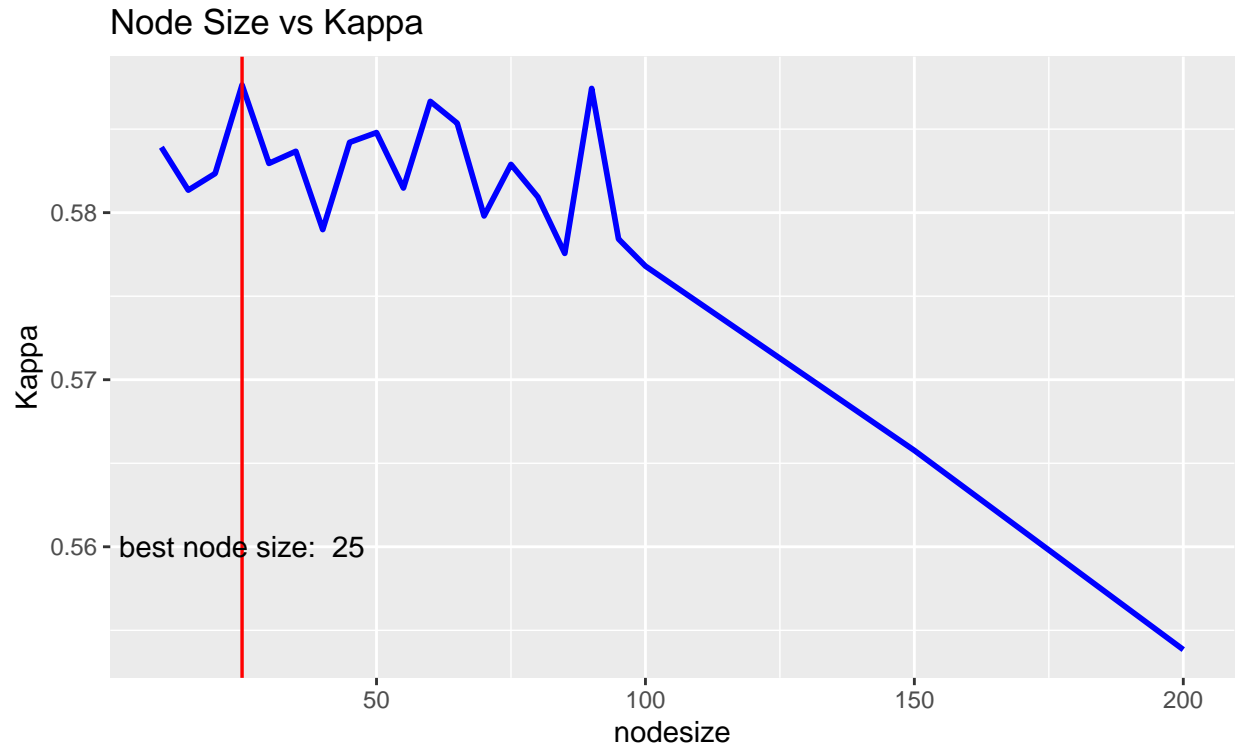
The following shows the cross validation results for model 2. Even though the best mtry from model 2 is different from the default one selected by caret in model 1, model 2 did not result in any improvement over model 1; ROC is still at 90% with both sensitivity and specificity remaining the same. In fact, the ROC curves for both models look identical.



RandomForest Model 3. Attempt to tune randomForest nodesize parameter. The final two models involved attempting to tune randomforest mtry parameter. In model 1, caret selected the mtry;

model 2 tuned the mtry from model 1.

In model 3, mtry from model 2 is used and held it constant to find the best nodesize. Different models were fitted across different values of nodesize and the one with best Kappa was selected. The plot below shows the values of Kappa across different values for nodesize in model 3.



Models 3 suggests the number 25 as an ideal value for node size. The following table shows the performance statistics for all three models fitted on the test set, including the best model from knn.

Table 14: Model performance on Test Set

	accuracy	kappa	sensitivity	specificity	k	mtry	nodesize
knn_best_model	0.8489407	0.5574877	0.5885350	0.9316343	23	NA	NA
rf_model.1 (using caret defaults)	0.8397298	0.5234003	0.5490446	0.9320388	NA	NA	NA
rf_model.2 (find best mtry)	0.8541603	0.5737706	0.6025478	0.9340615	NA	15	NA
rf_model.3 (find best nodesize)	0.8566165	0.5813384	0.6089172	0.9352751	NA	15	25

Final model and final prediction on validation set. From table above, randomForest model 3 (*rf_model*) is the best performing model. Even though the performance between kNN and the last two models from randomForest are very close in terms of accuracy, it is clear that randomForest model 3 is a better model using Kappa as a metric. Therefore, I still selected randomForest model 3 as the final model to apply on the validation because. The selected model used mtry of **15** and nodesize of **25**.

```

predict_final <- predict(fit_3, adult_income_validation, type="class")
confusionMatrix_final <- confusionMatrix(predict_final,
                                           adult_income_validation[, "over50K"])

$positive
[1] "Y"

$table
      Reference
Prediction  Y      N
      Y  2324   761
      N  1522 11673

$overall
      Accuracy      Kappa  AccuracyLower  AccuracyUpper  AccuracyNull
      8.597666e-01  5.828914e-01  8.543378e-01  8.650665e-01  7.637592e-01
AccuracyPValue  McnemarPValue
      4.014586e-205  5.758806e-57

$byClass
      Sensitivity      Specificity      Pos Pred Value
      0.6042642      0.9387968      0.7533225
      Neg Pred Value      Precision      Recall
      0.8846533      0.7533225      0.6042642
      F1      Prevalence      Detection Rate
      0.6706103      0.2362408      0.1427518
Detection Prevalence  Balanced Accuracy
      0.1894963      0.7715305

$mode
[1] "sens_spec"

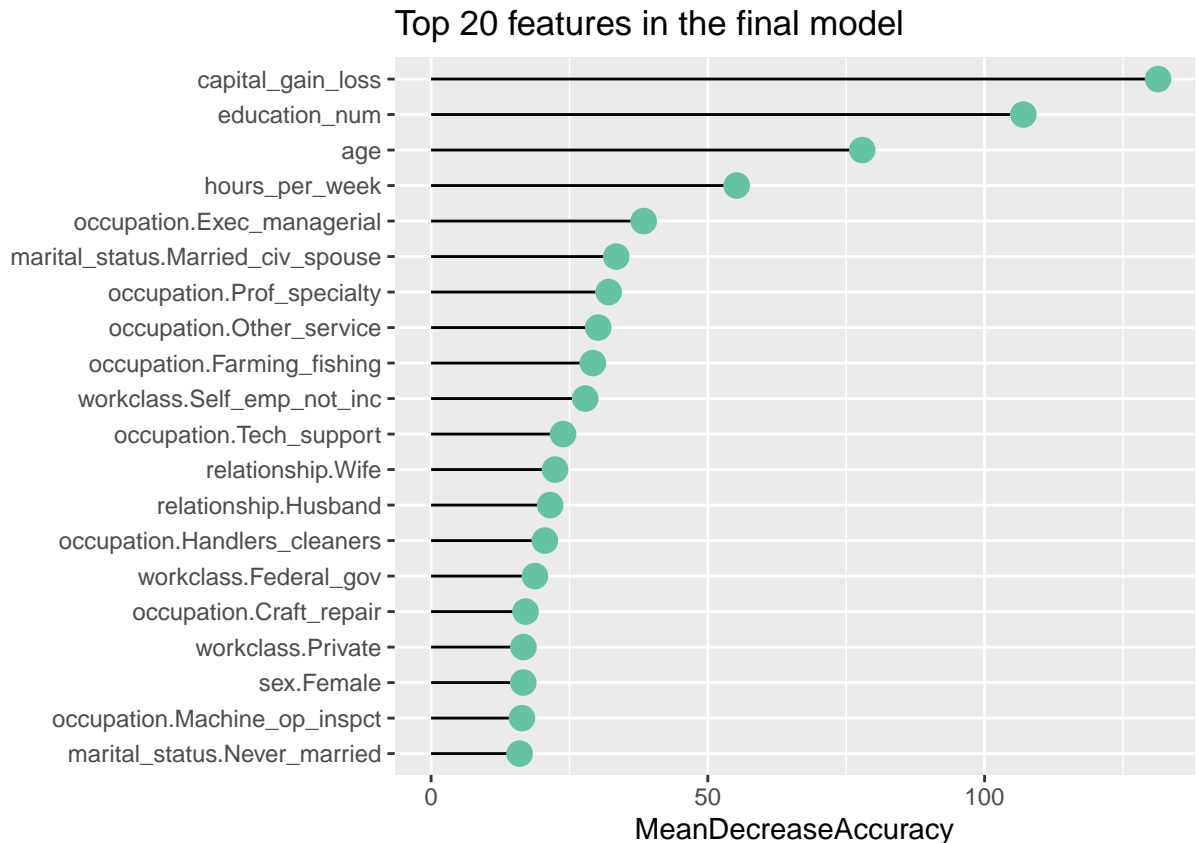
$dots
list()

attr("class")
[1] "confusionMatrix"

```

The final confusion matrix shows an Accuracy of **86%**, Kappa of *0.58*, and Sensitivity of **0.6**.

Important Variables. The randomForest model ranks the features (variables) used in the classification. The chart below shows the top 20 most important features in classifying whether a given income is >50K or not. From the chart, it can be seen that the capital_gain/loss variable is far the most important variable in the model. This is followed by education_num (number of years of education), and hour_per_week.



Conclusion

In this project, kNN and randomForest were used to predict whether an individual's income will be $\leq 50K$ or $> 50K$. Both algorithms performed very competitively, with randomForest winning based on Kappa performance. The final model on the validation set yielded an overall accuracy of 86%. Even with the low prevalence of the positive class, the model is able to correctly predict the positive class 60% of the time, which is pretty good. The model also showed that capital gain or loss is the most important variable in the classification of whether an individual's income is $\leq 50K$ or $> 50K$. The top four variables in the model for predicting the outcome are *capital_gain_loss*, *education_num*, *age*, and *hour_per_week*.

The project, for the sake of time, focused on two algorithms (kNN and randomForest). Although the final model's performance is fairly good, an improvement may have been achieved by considering more models and using ensemble methods. However, the skills acquired through this project is by far the most valuable to me.

REFERENCES

1. Irizarry, Rafael A., (2021), Introduction to Data Science Data Analysis and Prediction Algorithms with R, URL: <https://rafalab.github.io/dsbook/>
2. James, G., Witten, D., Hastie, T., Tibshirani, R., (2017), Introduction of Statistical Learning with Applications in R, “Springer Text in Statistics”.
3. Nwanganga, F & Chapple, Mike, (2020), Practical Machine Learning in R, “Wiley”.
4. Zhu, Hao. (2020), Create Awesome LaTeX Table with knitr::kable and kableExtra, URL: https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_pdf.pdf
5. U.S. Bureau of Labor Statistics, Standard Occupational Classification and Coding Structure, https://www.bls.gov/soc/2018/soc_2018_class_and_coding_structure.pdf
6. U.S. Bureau of Labor Statistics, Standard Occupational Classification Manual, https://www.bls.gov/soc/2018/soc_2018_manual.pdf
7. Wickham, H & Golemund, G; R for Data Science - Visualize, Model, Transform, Tidy, and Import Data, <https://r4ds.had.co.nz/>
8. Zhu, Hao, (2021), Create Awesome HTML Table with knitr::kable and kableExtra, URL: https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html