Predictive Models of Adult Income in the US Using Census Data from 1994

Bardh Kukaj

10/02/2024

Contents

1 Introduction

This report is written for the course "HarvardX PH125.9x - Data Science: Capstone," which is part of the HarvardX Professional Certificate in Data Science Program.

This report utilizes data from the U.S. Adult Income Census of 1994 to create a model that predicts whether an adult earns over or under 50k USD annually, based on 14 independent variables. The data, obtained from Kaggle and stored in the author's GitHub repository for this project (accessible in the R script or RMD file), is publicly shareable under the "CC0: Public Domain" license on Kaggle.

The objective of this report is to develop a model that achieves the highest possible prediction accuracy by utilizing various models, techniques, and algorithms. Another objective of this report is to investigate the determinants of earning annual income of over 50k USD.

To meet these objectives, six different models were developed, including two linear models (using logistic regression) and four non-linear models (two decision trees and two random forest algorithms).

The random forest model, using default settings, yielded the most optimal results. The models were compared using two metrics: Overall Accuracy and F1 Score. Overall Accuracy measures the model's effectiveness across all classes, while the F1 Score balances precision and recall, leading to a model that optimally predicts all classes by accounting for differences between them. The default random forest model achieved an overall accuracy of 0.87 and an F1 score of 0.91.

The report is structured as follows. The next section, "Analysis and Methodology," details the rationale behind selecting each model and technique. It begins with a discussion of all variables, followed by their visualization in relation to the dependent variable, income. The subsequent section focuses on the results, including explanations of the models used and the Overall Accuracy and F1 results. It is worth noting that the models created using logistic regression underwent a Likelihood Ratio Test, which revealed a highly significant p-value, rejecting the null hypothesis that the baseline (or reduced) model fits the data better than the full model. This section also briefly discusses the estimates of the reduced and full logistic regression models, which are attached as annexes at the end of this report. The concluding section, "Concluding Remarks," summarizes the findings, discusses the report's limitations, and offers recommendations for future work.

It should be noted that three files in total are submitted as part of this project: the R script, the RMD file, and the PDF file generated by the RMD file. To keep the PDF file succinct and avoid overwhelming the reader with details, most of the code found in the first two files has not been included in the report (pdf file).

2 Analysis and methodology

2.1 Exploring and cleaning data

After downloading and installing the necessary packages for this project, the data is retrieved from the GitHub repository link provided in the R script and the RMD file. The data is then cleaned by removing instances of "NAs" and "?" that are present in the dataset. Initially, there are 32,561 observations; after cleaning, 30,162 remain.

There are a total of 15 variables, which are listed and briefly explained below.

- age
- work class - type of work the respondents are engaged, e.g. private, federal, etc.
- fnlwgt - or final weight is an estimate of the number of people in the US each observation (or respondent) represents.
- education - level of education.
- education. num - years of education.
- marital.status - categories include: "Divorced", "Married-AF-spouse", "Married-civ-spouse", "Marriedspouse- absent", and "Never-married", "Separated", and "Widowed".
- occupation - type of occupation.
- relationship - categories include: "Husband", "Not-in-family", "Other-relative", "Own-child", "Unmarried", and "Wife".
- race - categories include: "Amer-Indian-Eskimo", "Asian-Pac-Islander", "Black", "Other", and "White".
- sex - male and female.
- capital. gain - whether the respondent has capital gains.
- capital.loss - whether the respondent has capital loss.
- native.country - includes 41 countries.
- income - categories include: over 50k USD, or under 50k USD (annually).

The dataset contains two types of variables: nominal categorical variables and discrete numerical variables. Before partitioning the dataset into training and test sets, with a ratio of 8:2, certain variables are designated as categorical (or factors, in R programming language terminology). These variables include "workclass", "sex", "race", "marital.status", "occupation", "native.country", "education", "relationship", and "income". The remaining variables are specified as integers. Given that the dependent variable, income, is also categorical, the selection of tools and algorithms is made accordingly. The chosen methods include logistic regression, decision trees, and random forest algorithms.

It should be noted that for machine learning purposes, this dataset is not considered large (but neither it is too small), therefore we use the 8:2 ratio which is a common standard in machine learning.

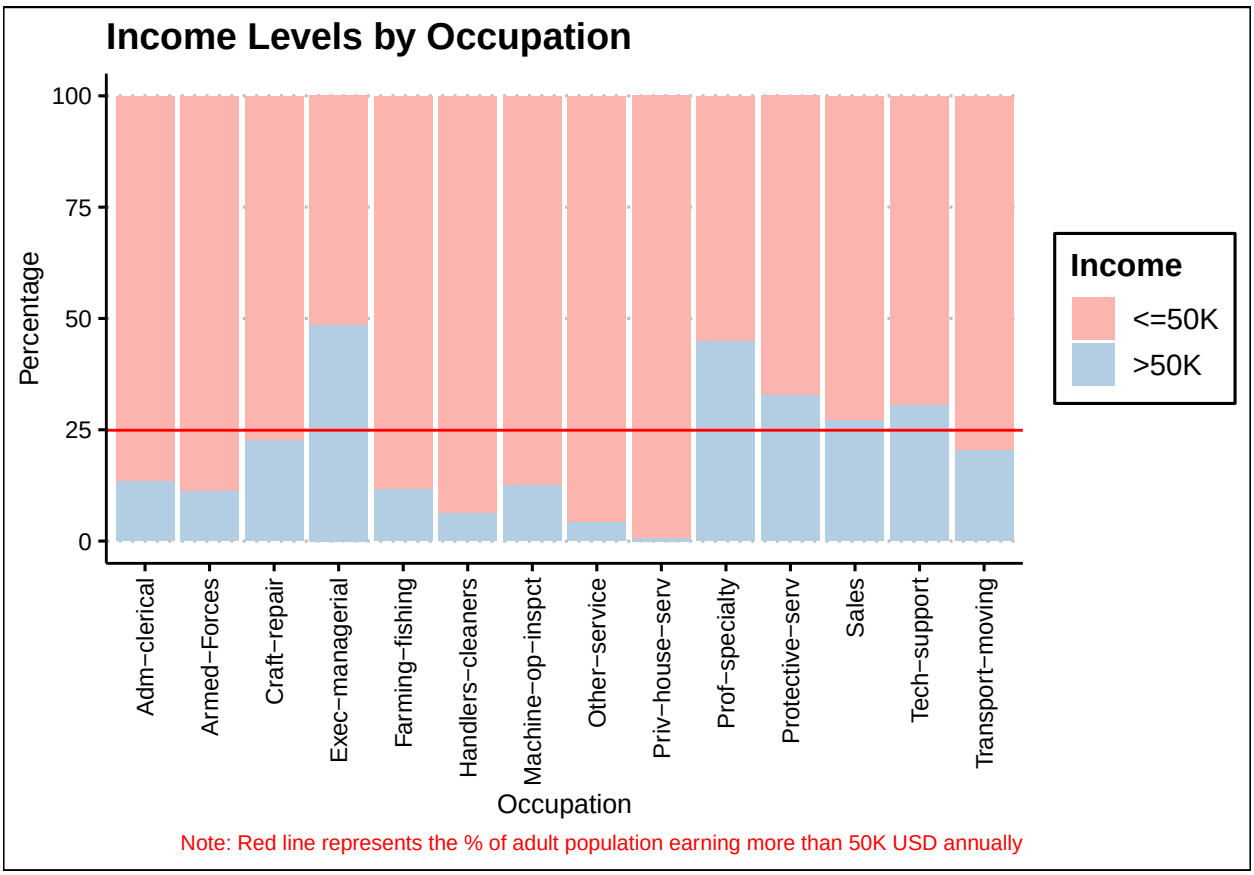
Lastly, since the dataset is used primarily for the purpose of identifying a model that optimally fits the data, rather than for making inferences about the total population, I will not adjust the numerical variables to reflect exact population numbers. However, I will include it as an independent variable in all models, except for the first baseline model.

2.2 Visualising the relationship between independent variables with the dependent variable

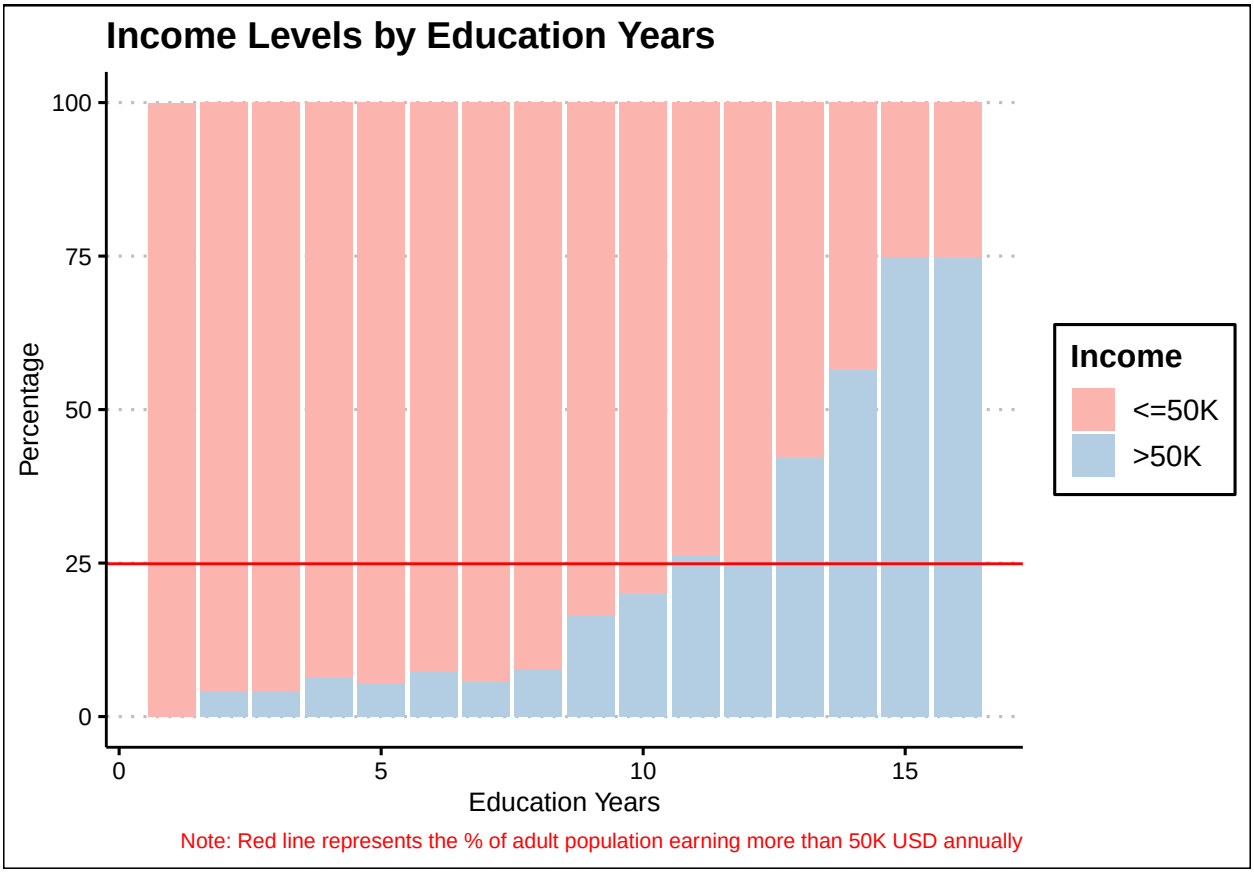
The initial set of variables selected for visualization are those that, intuitively, have a strong economic rationale for influencing income levels. These variables include "occupation," "education," "workclass," "capital.gain," "capital.loss," and "hours.per.week." It's important to note that including "education.num" alongside "education" can lead to multicollinearity, as both variables essentially convey the same information but in different forms. Since "education" provides a more statistically significant estimate than

"education.num," only the former is included in the analysis to avoid redundancy and potential distortion in the model's results.

The analysis also calculates the average income level across all respondents, revealing that approximately 25 percent of them earn more than 50K USD annually. This insight is represented visually as a red line in all the following graphs, providing a clear benchmark against which to compare the income distribution across different categories and variables.

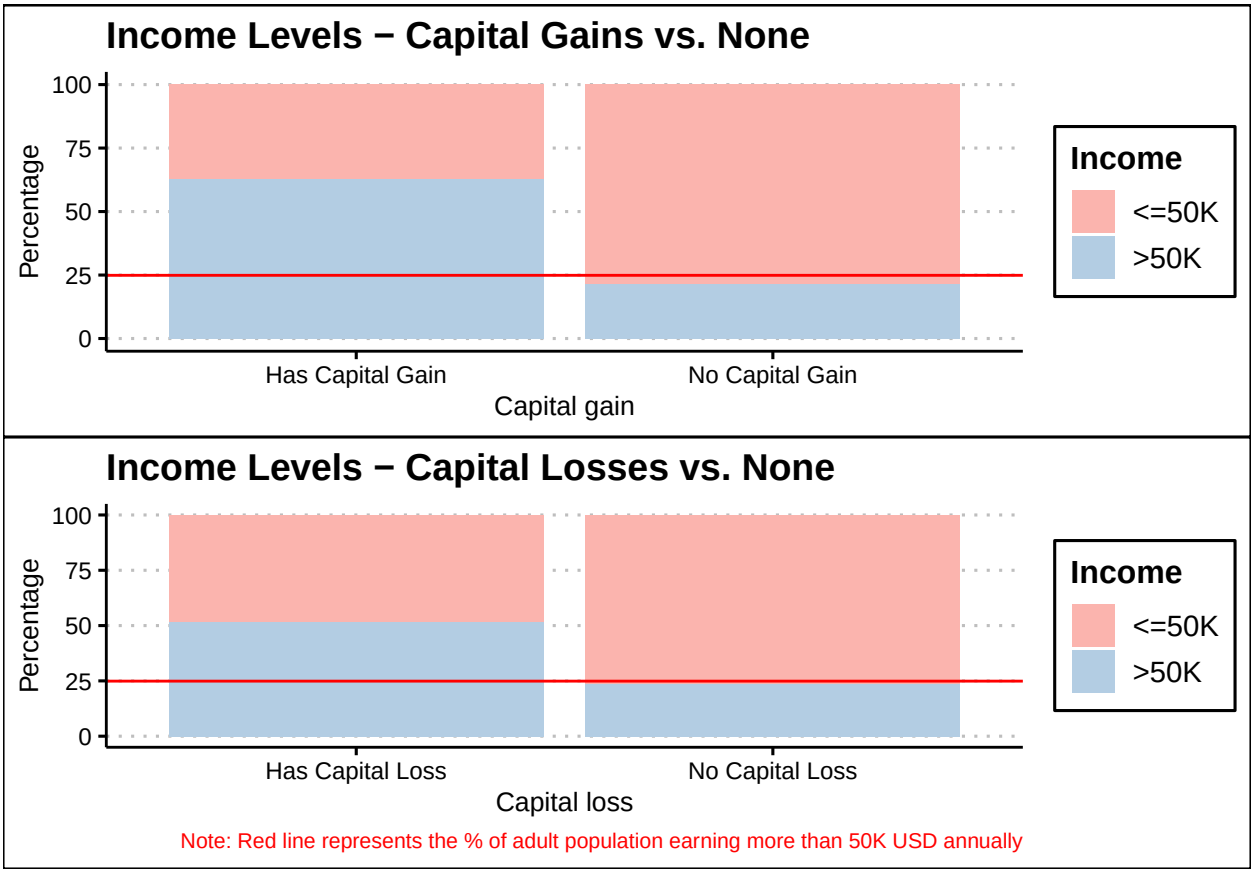


The graph above shows that occupations such as "Sales" and "Craft Repair" have average incomes around the 25 percent mark, indicating that approximately a quarter of individuals in these occupations earn more than 50K USD annually. On the other hand, "Executives" and professionals with "Specialties" significantly outperform the average, with about 50 percent of individuals in these categories surpassing the 50K USD annual income threshold. Conversely, occupations such as "Handlers/Cleaners" and those involved in "Private House Services" are at the lower end of the income spectrum, earning the least.

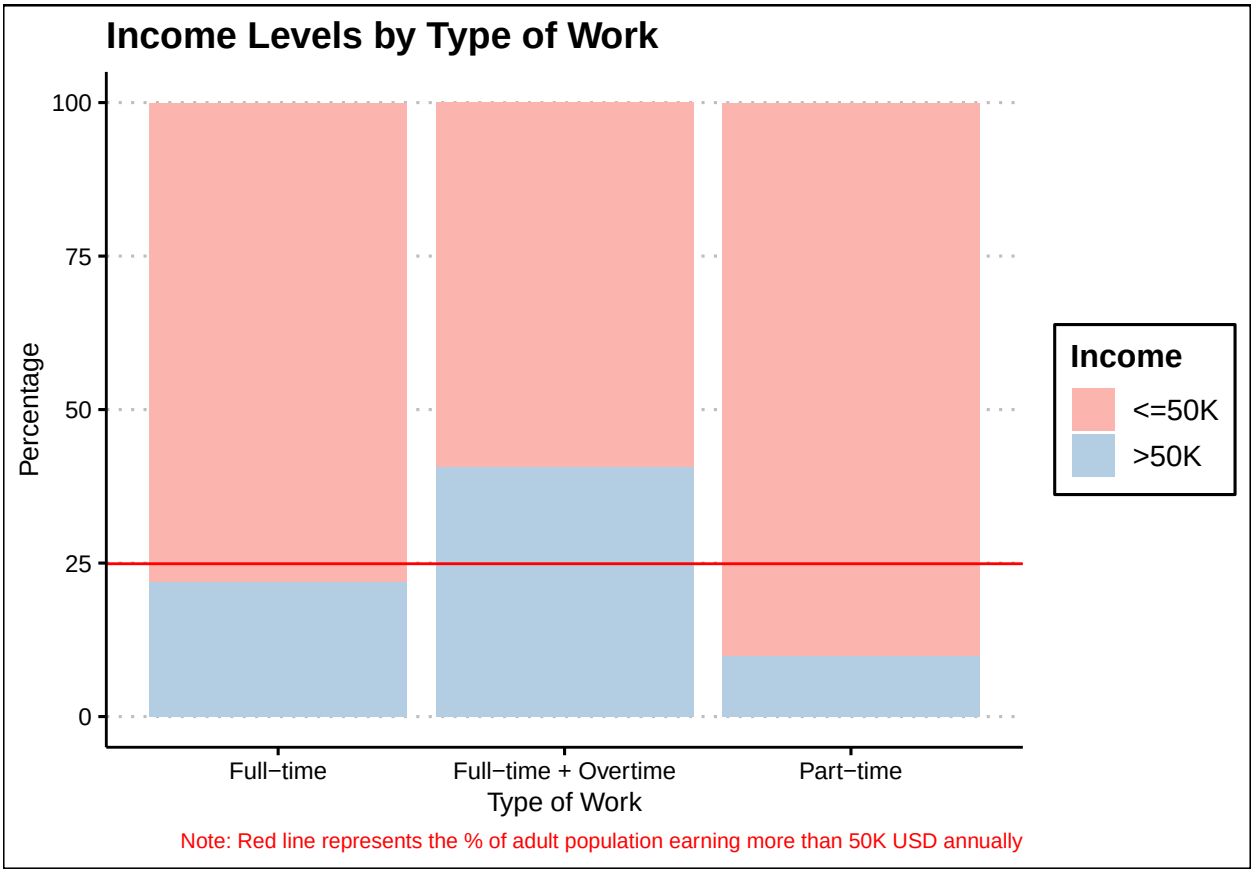


The impact of education on income appears relatively stable up to nine years of education, indicating minimal variation in income levels within this range. Beyond nine years of education there is a noticeable increase in income. This upward trend continues until one reaches 15 years of education, after which the income levels off, showing similar income yields for individuals with 15 and 16 years of education.

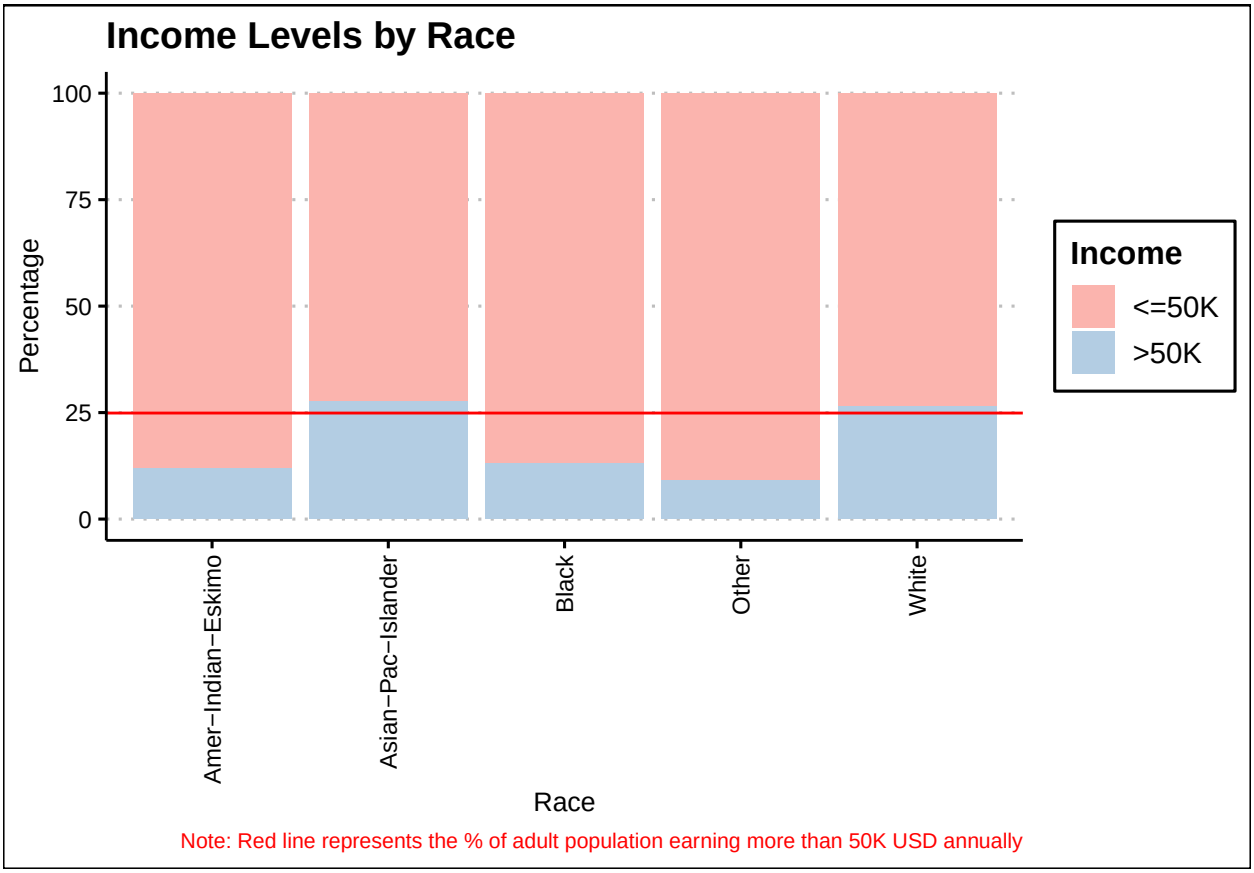
When analyzing the distribution of income across various work classes, it can be noticed that the majority align closely with the overall mean, which indicates that about 25 percent of the adult population earns more than 50K USD annually. Notably, individuals employed by the federal government and, more significantly, those who are self-employed, stand out compared to this trend. Approximately 35 percent of federal government workers and over 50 percent of self-employed individuals report earning more than 50K USD annually.



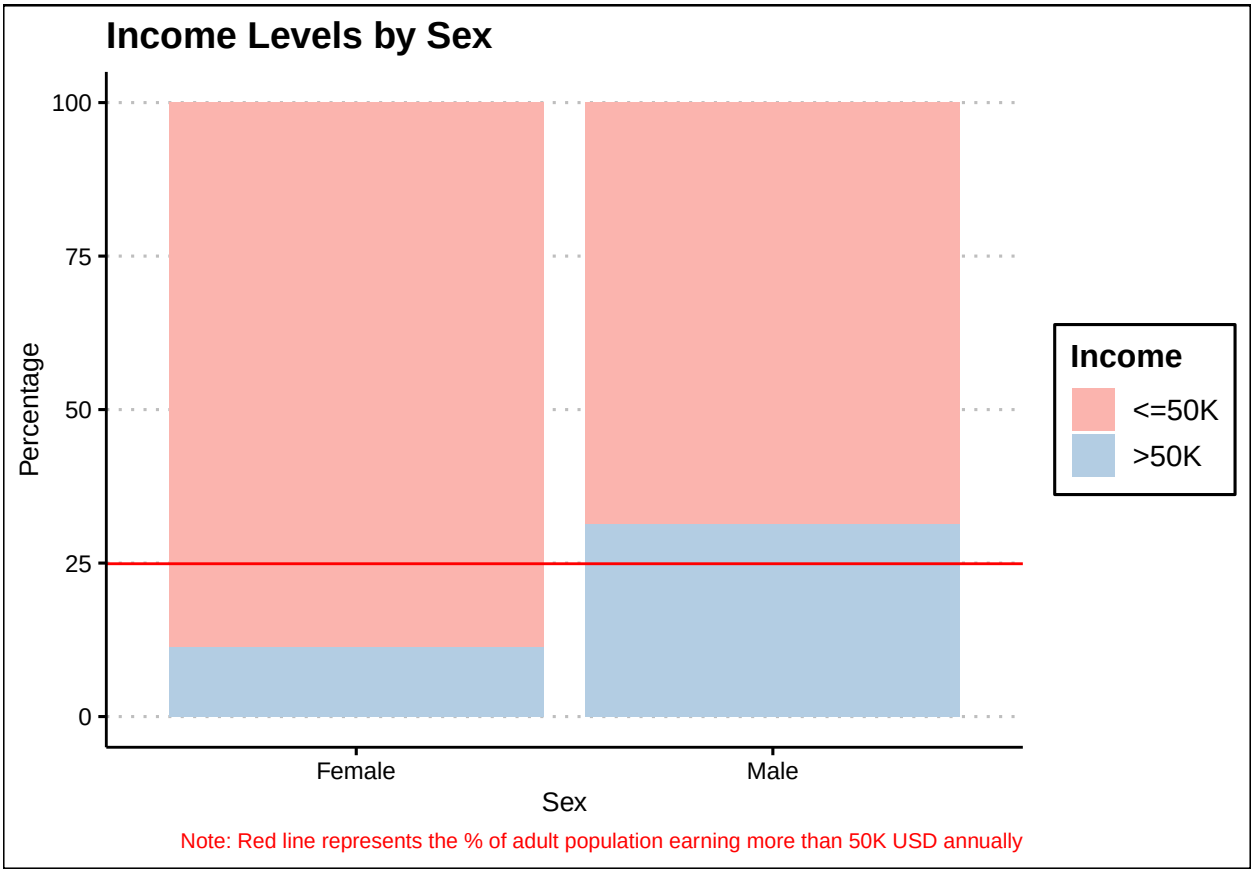
It is reasonable to assume that individuals reporting either capital gains or losses are generally in higher income brackets compared to those who do not report such financial activities. This assumption is supported by the graph above. Specifically, the bar plot reveals that approximately 63 percent of individuals with capital gains earn more than 50K USD annually, a higher percentage compared to over 50 percent of those reporting capital losses. This indicates a clear correlation between the presence of capital gains or losses and higher income levels.



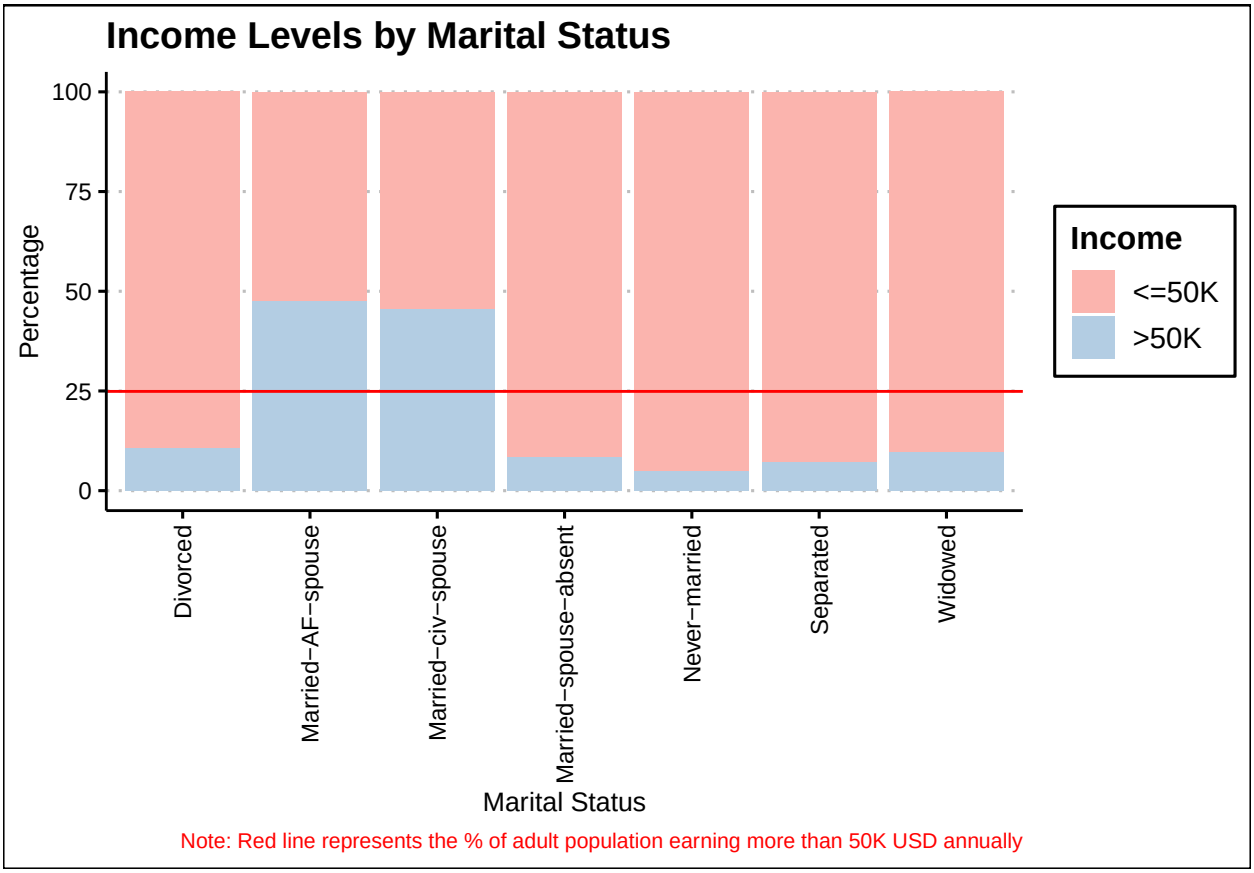
For the purpose of visual analysis, data on work hours per week have been segmented into three categories: part-time (less than 40 hours per week), full-time (exactly 40 hours per week), and full-time with overtime (more than 40 hours per week). The visual representation reveals an interesting trend: on average, individuals working exactly 40 hours per week (full-time workers) earn less than the average income observed across all respondents.



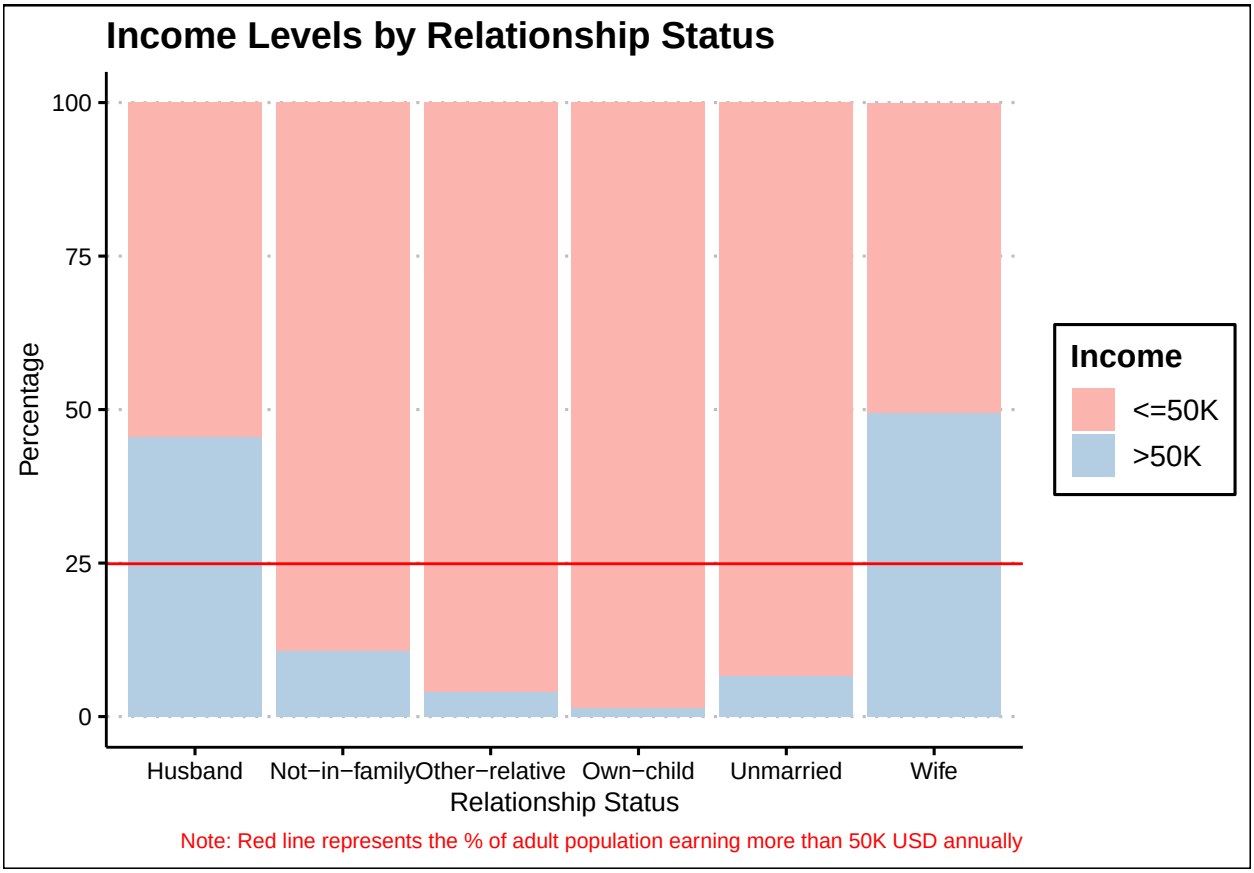
When analyzing income levels by race, it can be noticed that, on average, individuals identified as Amer-Indian-Eskimo have the highest earnings. As previously mentioned, the data has not been adjusted for the final weight variable, which could explain the unexpected result when attempting to generalize these findings to the total population. White respondents appear to earn slightly above the average, whereas individuals from other racial groups earn significantly below the average.



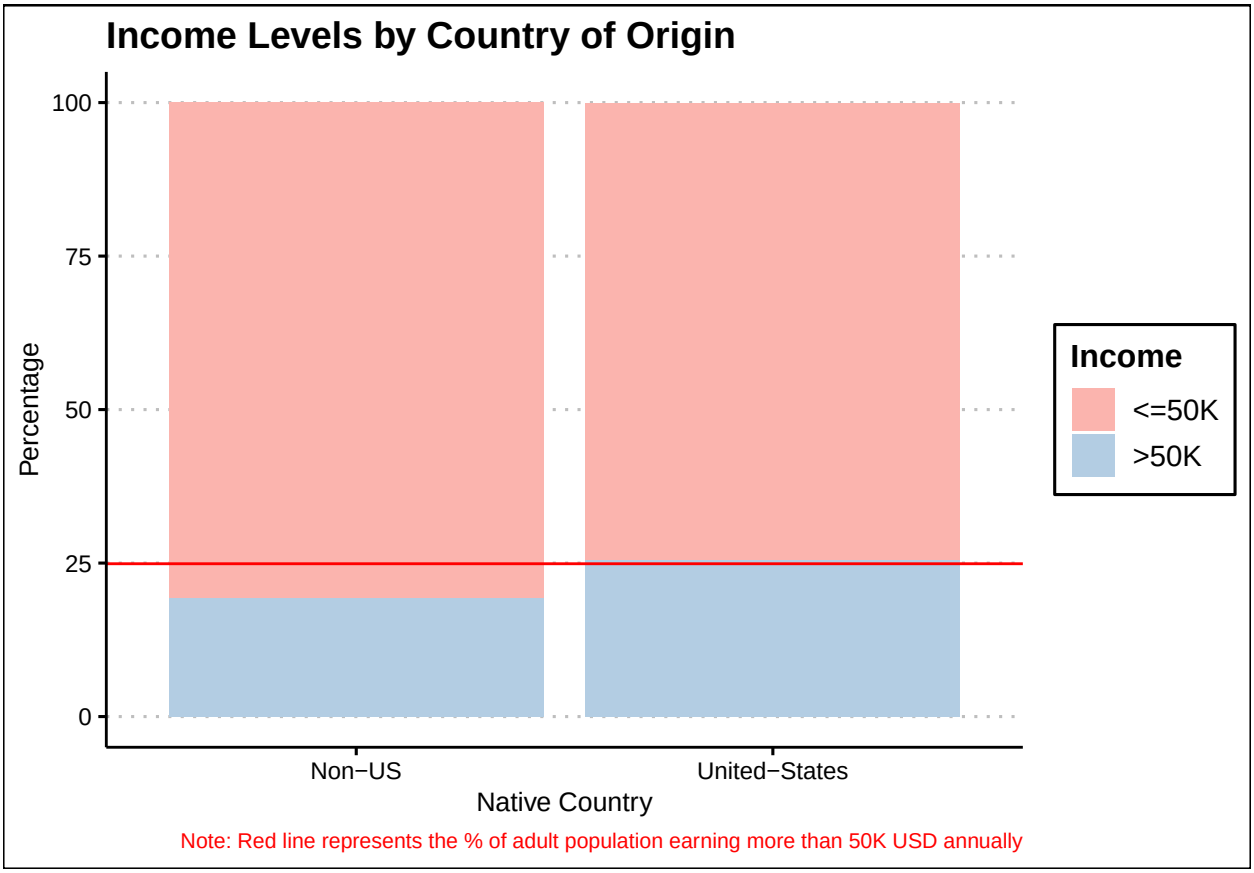
The analysis of the gender pay gap, using data from 1994, confirms that men, on average, earn more than women. Specifically, men's earnings exceed the overall average, while women's earnings are less than half of this average. This stark disparity highlights the significant income gap between genders as observed in the dataset from that period.



Those married with an armed forces (AF spouse) or with a civilian (civ spouse) by far earn more than the rest, with the former earning the most. Those who never married on average have the least income.



The results of those being married (with a present spouse) are replicated also in the following graph, where those that identify either as a husband or a wife generally earn more than the rest.



The data for the graph above have been pre-processed into two categories: United-States and Non-US. Originally, there are around 41 native countries to which the respondents belong. Those whose native country is United-States generally earn more than the average, and effectively the rest.

The visualized relationships between the independent variables and the dependent variable (income) highlight that income is indeed distributed differently among different groups (e.g., races, sexes, or relationships among others). While this might be intuitive, the evidence presented so far is insufficient to conclude that there are biases in terms of how income was distributed in the US in 1994. The next section, besides trying to find an optimal model that explains best the variations in the dependent variable, will try to shed light to what extent economic factors determine whether an individual makes it past the 50k USD threshold, and how biases affect these dynamics.

2.3 Specification of models

A total of six models have been developed with the aim of increasing the overall accuracy and F1 score, so that the predicted values fit the actual data as best as possible. Two models have been created using each of the following techniques: logistic regression, decision trees, and random forest.

This section will merely outline the model used. The results will be shown and discussed in the following section.

2.3.1 Logistic Regression Baseline Model

The first model, the baseline logistic regression model, consists of variables that intuitively have an economic reasoning for determining the income level. These are occupation, education, workclass, capital.gain, capital.loss, and hours.per.week. As mentioned earlier, adding education.num leads to multicollinearity, as

education is already added to the equation. The estimate of education is more statistically significant than education.num, therefore the latter is not added.

As the dependent variable is categorical (a factor), logistic regression is a technique that is suitable as it measures the probability of of an event occurring - in this case, the probability whether a respondent receives more or less than 50K USD annually.

The mathematical formulation of the model is shown below.

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 \cdot \text{Workclass} + \beta_2 \cdot \text{Education} + \beta_3 \cdot \text{Occupation} + \beta_4 \cdot \text{Capital Gain} + \beta_5 \cdot \text{Capital Loss} + \beta_6 \cdot \text{Hours per Week}$$

2.3.2 Logistic Regression Full Model

As the title shows, the same technique is used also here, but different from the previous model, in this model all variables are added as independent variables. Adding the education.num variable does not make a difference, as the package recognizes multicollinearity and drops it from the equation.

Another reason for creating another model using logistic regression, besides seeking to optimize overall accuracy and F1 score, is to run a statistics test (likelihood ratio test) that is used to compare the goodness-of-fit of two models, and is often used in logistic regression models.

2.3.3 Decision Tree Default Settings Model

Decision trees are recommended for handling both categorical and numerical variables, allowing them to capture non-linear relationship. This model includes all the independent variables and it uses the default settings provided by the "caret" package.

It can be easily understood and interpreted by the reader as it can be plotted as a tree-like model of decision and the possible consequences of each decision.

2.3.4 Controlled Decision Tree Model

This model is developed using a decision tree algorithm, which is further optimized by tuning the complexity parameter (cp) through cross-validation. A 10-fold cross-validation is used in this model. This means that the dataset is divided into 10 parts, and the model is trained and tested 10 times. The cp, or complexity parameter, controls the size of the decision tree and ensures there is no overfitting of the model. Based on the most optimal cross-validation results, as the model is trained numerous times with different cp values, the most optimal model is then selected. This model uses for the cp all the values from 0.01 to 0.1, in an increment of 0.01.

2.3.5 Random Forest Default Settings Model

A random forest is an ensemble machine learning algorithm that combines the outputs of multiple decision trees to make more accurate predictions. Essentially, it aggregates the predictions from numerous decision trees to reduce the risk of overfitting, which is common when using a single decision tree. This approach generally results in more reliable and robust estimates.

In this model, all independent variables are utilized to explain variations in the income variable, employing the default settings of the "caret" package. The use of default settings simplifies the model-building process, allowing for an efficient evaluation of the predictive power of the independent variables on income, while leveraging the strength of the random forest algorithm in handling complex data structures and relationships.

2.3.6 Random Forest Tuned Model

Similar to the controlled decision tree model, this model as well uses a 10-fold cross-validation to find the optimal value of the "mtry". The "mtry"is a parameter used in machine learning that specifies the number of features (or variables) randomly sampled at each decision point of the random forest. This model uses four different values for the mtry, namely: 1, 3, 5, and 7. 16

3 Results

Attempts to obtain predicted values that fit better the actual data were successful for the most part. The table below shows that the the most performing model is the random forest with default settings from the "caret" package, while the decision tree models did not yield the desired results. All models were compared by their "Overall Accuracy" and "F1 score".

Table 1: Summary of Results

method	Overall_Accuracy	F1_Score
Baseline Logistic Regression	0.81618	0.88575
Logistic Regression + biases	0.85016	0.90300
Decision tree - Default Settings	0.82695	0.89116
Controlled Decision Tree	0.83889	0.89743
Random Forest - Default Settings	0.86275	0.91053
Random Forest - Tuned	0.85579	0.90751

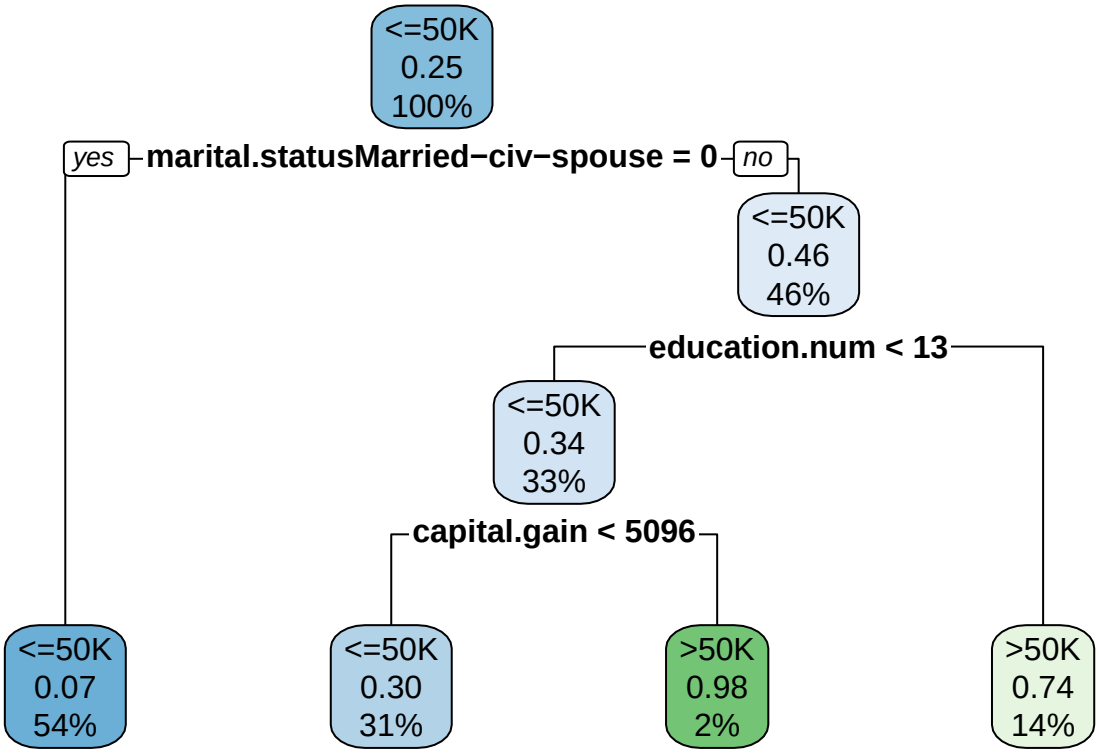
The improvement from the first to the second model (reduced or baseline logistic regression model to the full regression model) is intuitive as in the latter all the independent variables are added, compared to the former which includes only variable that consist of an economic reasoning for explaining the variations in income. Annex 1 and Annex 2 show a summary of the models, namely the estimates, standard errors, z-value, and the p-value. Statistical significance is also showed with asterisks (more indicate higher statistical significance). The estimates show the log odds of earning more than 50K USD annually for each independent variable, while holding the other variables constant. A positive estimate indicates a higher likelihood of earning more than 50K USD annually, and vice versa. It should be highlighted that adding all the variables as predictors in the second logistic regression model does not only increase the "Overall Accuracy" and "F1 Score", it also shifts the statistical significance of some variables, e.g., "education7th-8th" in the second model is statistically significant, while in the first it is not; the same case is with "workclassSelf-emp-inc'".

A likelihood ratio test was performed for the reduced and the full logistic regression models. It showed that the latter model provides a significantly better fit for the data, as the p-value is 2.2e-16. This shows that the additional variables added in the full model improve the explaining of the variance in the data.

The results of the full logistic regression model (shown in Annex 2) are quite intuitive. For instance, the age variable (which is highly statistically significant) shows that with each added year of age (which can be interpreted also as work experience) the odds of earning more than 50k USD annually increase by 2.6 percent. In terms of education, having a bachelor's degree increases the odds of belonging to the threshold beyond 50k USD annually by almost 7 times compared to the reference category, which is dropped in the results - that is having only the 10th grade finished. However, biases can be noticed in the results as well, and the main one, which is highly statistically significant, is being a male. Male respondents' odds of earning over 50k USD annually are 2.32 times higher compared to females. All these interpretations mean that all the other variables are held constant. The sex variable shows that there was evidence of gender pay gap in the US in 1994. Such results, unfortunately, are evident worldwide.

Expecting improvement in results when moving to decision tree models from logistic regression models might be intuitive as we shift from linear models (logistic regression) to non-linear models (decision trees) which capture the variations better. However, that is not the case with this data. An explanation to this might be that the relationship between the independent variables and the dependent variable is generally linear, therefore the decision tree models do not yield better result. This is a potential venue for other students/researchers to explore.

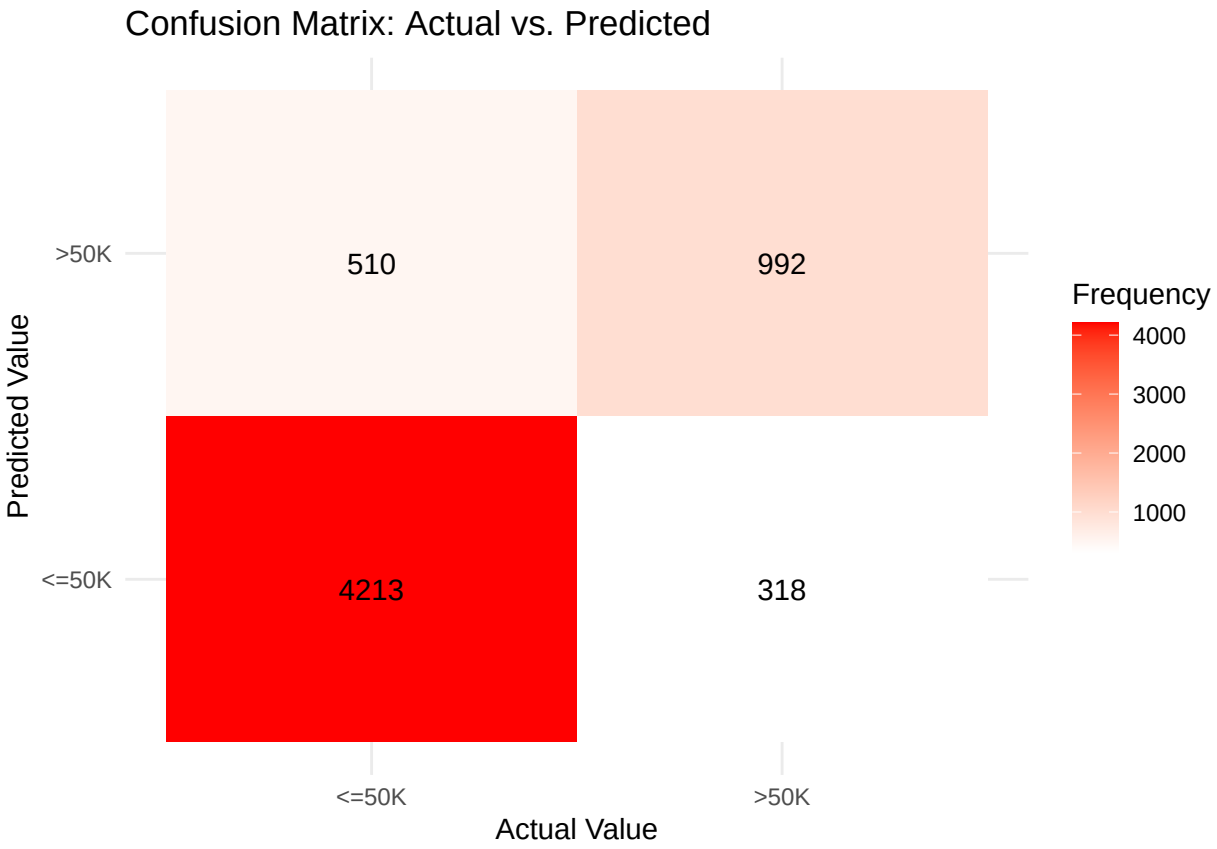
Below is shown the decision tree plotted from the first decision tree model. The first figure shows the predicted class (or the categorical dependent variable). The second figure shows the probability of the observations in this node that belong to the predicted class. The third figure shows the percentage of observations that fall in that node.



Lastly, random forest, an ensemble method, which handles better class imbalances (as in this case there are far more observations falling under the category of "<=50K", than ">50K") provides the most optimal results. The tuned random forest model does not yield the desired results, mainly due to limitations of computational power of the machine used for this project. This is different from the controlled decision tree model which provides better results than the decision tree model with default settings.

As the fifth model, the random forest model with default settings, yields the most optimal results, below are plotted the actual values vs. the predicted values.

A heatmap is used to plot the values. Note that due to imbalances among classes, "<=50K" vs. ">50K", the bottom left tile is a dark red due to higher frequency of observations (around 75 percent of respondents earn less than 50K USD annually). The model wrongly predicts more cases of those earning more than 50K, when they actually earn the opposite, than the other way around.


Confusion Matrix: Actual vs. Predicted
>50K
<=50K
Predicted Value
510
992
4213
318
<=50K
>50K
Actual Value
Frequency
4000
3000
2000
1000

4 Conclusion and recommendations

The purpose of this project was to develop a model that achieves the highest accuracy prediction accuracy by utilizing different models and different techniques and algorithms.

Adding explanatory variables has shown that biases were present in the 1994 era when determining the income of an individual. A variable that is a source of bias, that is highly statistically significant, is sex.

Several iterations of data pre-processing have been conducted to try to achieve better results in terms of "Overall Accuracy" and "F1 score". Several categories of variables such as native.country or education have been grouped to try and achieve higher statistical significance of any, but none provided better results than leaving them as they are. After all, they do provide more complexity, and this might be the reason why it has hard to achieve better results.

Using different techniques/algorithms has proved useful. Logistic regression, decision trees, and random forest are all used with categorical data. Following the theoretical explanation that ensemble methods (random forest in this case) handle better class imbalances (as there are more observations of "<=50K", than ">50K") helped achieve the most optimal results for this project. However, computational power of the machine used for this project was insufficient to achieve better results with a tuned random forest model, despite trying different values for cross-validation and mtry, and ways of choosing the mtry.

A recommendation for other students and researchers is to explore and determine the shifts in statistical significance when comparing the reduced and the full logistic regression models. It could be due to better model specification (more explanatory variables explain better the variations in the dependent variable), or reduced multicollinearity among others. Comparing the results of more recent census data would be beneficial for the public in order to highlight the changes, if evident, in terms of income distribution and the presence of biases. Another recommendation is to adjust the results for the final weight variable in order to infer about the whole population, and not merely the data set as this project did.

5 Annex 1 - Summary of the Baseline Logistic Regression Model

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0742  -0.6436  -0.4116   0.0000   2.9710
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -3.745e+00  2.024e-01 -18.503  < 2e-16 ***
## `workclassLocal-gov`         -7.457e-01  1.101e-01  -6.771 1.28e-11 ***
## workclassPrivate             -6.760e-01  9.096e-02  -7.432 1.07e-13 ***
## `workclassSelf-emp-inc`       5.012e-02  1.225e-01   0.409  0.68236
## `workclassSelf-emp-not-inc`  -6.821e-01  1.087e-01  -6.277 3.46e-10 ***
## `workclassState-gov`         -8.479e-01  1.223e-01  -6.932 4.14e-12 ***
## `workclassWithout-pay`       -1.089e+01  8.926e+01  -0.122  0.90286
## education11th                -1.620e-01  2.271e-01  -0.713  0.47583
## education12th                 9.964e-02  2.924e-01   0.341  0.73331
## `education1st-4th`           -3.323e-01  5.215e-01  -0.637  0.52406
## `education5th-6th`           -2.768e-01  3.617e-01  -0.765  0.44415
## `education7th-8th`           -3.402e-01  2.715e-01  -1.253  0.21009
## education9th                 -1.895e-01  2.961e-01  -0.640  0.52215
## `educationAssoc-acdm`         1.067e+00  1.879e-01   5.678 1.36e-08 ***
## `educationAssoc-voc`          1.098e+00  1.831e-01   6.000 1.97e-09 ***
## educationBachelors            1.604e+00  1.707e-01   9.398  < 2e-16 ***
## educationDoctorate            2.801e+00  2.249e-01  12.458  < 2e-16 ***
## `educationHS-grad`            6.993e-01  1.675e-01   4.176 2.96e-05 ***
## educationMasters              2.055e+00  1.802e-01  11.404  < 2e-16 ***
## educationPreschool           -2.152e+01  1.942e+02  -0.111  0.91175
## `educationProf-school`        2.710e+00  2.124e-01  12.759  < 2e-16 ***
## `educationSome-college`       8.840e-01  1.693e-01   5.221 1.78e-07 ***
## `occupationArmed-Forces`     -1.131e+00  1.164e+00  -0.971  0.33131
## `occupationCraft-repair`      7.297e-01  7.653e-02   9.535  < 2e-16 ***
## `occupationExec-managerial`   1.219e+00  7.286e-02  16.731  < 2e-16 ***
## `occupationFarming-fishing`  -3.013e-01  1.430e-01  -2.107  0.03509 *
## `occupationHandlers-cleaners` -4.814e-01  1.472e-01  -3.270  0.00108 **
## `occupationMachine-op-inspct`  2.317e-01  1.022e-01   2.266  0.02345 *
## `occupationOther-service`    -8.366e-01  1.197e-01  -6.991 2.73e-12 ***
## `occupationPriv-house-serv`  -4.548e+00  2.380e+00  -1.911  0.05602 .
## `occupationProf-specialty`    7.133e-01  7.808e-02   9.136  < 2e-16 ***
## `occupationProtective-serv`   1.119e+00  1.223e-01   9.150  < 2e-16 ***
## occupationSales               6.116e-01  7.791e-02   7.850 4.16e-15 ***
## `occupationTech-support`      8.280e-01  1.058e-01   7.824 5.10e-15 ***
## `occupationTransport-moving`  6.415e-01  9.759e-02   6.573 4.92e-11 ***
## capital.gain                  3.371e-04  1.165e-05  28.950  < 2e-16 ***
## capital.loss                  7.196e-04  3.724e-05  19.322  < 2e-16 ***
## hours.per.week                3.240e-02  1.643e-03  19.723  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 27079  on 24128  degrees of freedom
## Residual deviance: 19964  on 24091  degrees of freedom
## AIC: 20040
##
## Number of Fisher Scoring iterations: 11
```

6 Annex 2 - Summary of the Full Logistic Regression Model

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1155  -0.5159  -0.1900   0.0000   3.7565
##
## Coefficients: (2 not defined because of singularities)
##                                          Estimate Std. Error z value
## (Intercept)                             -5.831e+00  8.142e-01  -7.161
## age                                      2.591e-02  1.920e-03  13.493
## `workclassLocal-gov`                    -6.817e-01  1.256e-01  -5.428
## workclassPrivate                        -4.656e-01  1.041e-01  -4.470
## `workclassSelf-emp-inc`                 -2.925e-01  1.379e-01  -2.121
## `workclassSelf-emp-not-inc`             -9.745e-01  1.223e-01  -7.971
## `workclassState-gov`                    -8.291e-01  1.399e-01  -5.926
## `workclassWithout-pay`                  -1.431e+01  3.482e+02  -0.041
## fnlwgt                                   6.711e-07  1.964e-07   3.418
## education11th                            1.855e-01  2.364e-01   0.785
## education12th                            3.622e-01  3.171e-01   1.142
## `education1st-4th`                      -4.473e-01  5.550e-01  -0.806
## `education5th-6th`                      -3.115e-01  3.844e-01  -0.810
## `education7th-8th`                      -6.701e-01  2.821e-01  -2.376
## education9th                            -2.115e-01  3.086e-01  -0.685
## `educationAssoc-acdm`                    1.256e+00  2.008e-01   6.255
## `educationAssoc-voc`                     1.270e+00  1.946e-01   6.524
## educationBachelors                       1.931e+00  1.808e-01  10.683
## educationDoctorate                       3.012e+00  2.519e-01  11.957
## `educationHS-grad`                       7.914e-01  1.757e-01   4.505
## educationMasters                         2.270e+00  1.934e-01  11.738
## educationPreschool                      -1.941e+01  1.640e+02  -0.118
## `educationProf-school`                   2.913e+00  2.327e-01  12.522
## `educationSome-college`                  1.114e+00  1.783e-01   6.244
## education.num                                   NA         NA      NA
## `marital.statusMarried-AF-spouse`        2.840e+00  6.103e-01   4.653
## `marital.statusMarried-civ-spouse`       1.995e+00  3.058e-01   6.525
## `marital.statusMarried-spouse-absent`    2.697e-02  2.666e-01   0.101
## `marital.statusNever-married`           -5.256e-01  9.998e-02  -5.257
## marital.statusSeparated                 -1.089e-01  1.840e-01  -0.592
## marital.statusWidowed                    1.828e-01  1.765e-01   1.036
## `occupationArmed-Forces`                -1.136e+00  1.561e+00  -0.728
## `occupationCraft-repair`                 4.716e-02  9.042e-02   0.522
## `occupationExec-managerial`              7.927e-01  8.696e-02   9.116
## `occupationFarming-fishing`             -8.881e-01  1.555e-01  -5.712
## `occupationHandlers-cleaners`           -7.215e-01  1.625e-01  -4.439
## `occupationMachine-op-inspct`           -3.054e-01  1.156e-01  -2.642
## `occupationOther-service`               -8.045e-01  1.334e-01  -6.033
## `occupationPriv-house-serv`             -3.962e+00  1.888e+00  -2.098
## `occupationProf-specialty`               5.259e-01  9.257e-02   5.680
## `occupationProtective-serv`              5.748e-01  1.402e-01   4.099
## occupationSales                          2.354e-01  9.311e-02   2.529
```

```
## `occupationTech-support`                    6.251e-01  1.250e-01   5.001
## `occupationTransport-moving`               -7.442e-02  1.106e-01  -0.673
## `relationshipNot-in-family`                 3.404e-01  3.020e-01   1.127
## `relationshipOther-relative`               -5.349e-01  2.762e-01  -1.937
## `relationshipOwn-child`                    -8.012e-01  3.026e-01  -2.648
## relationshipUnmarried                       2.544e-01  3.196e-01   0.796
## relationshipWife                            1.292e+00  1.183e-01  10.917
## `raceAsian-Pac-Islander`                    5.581e-01  3.086e-01   1.808
## raceBlack                                   2.133e-01  2.559e-01   0.834
## raceOther                                   2.189e-03  4.104e-01   0.005
## raceWhite                                   3.620e-01  2.422e-01   1.495
## sexMale                                     8.415e-01  9.044e-02   9.305
## capital.gain                                3.200e-04  1.202e-05  26.614
## capital.loss                                6.548e-04  4.280e-05  15.298
## hours.per.week                              2.871e-02  1.909e-03  15.040
## native.countryCanada                       -1.053e+00  7.387e-01  -1.425
## native.countryChina                        -1.777e+00  7.452e-01  -2.385
## native.countryColumbia                     -3.256e+00  1.080e+00  -3.014
## native.countryCuba                         -9.868e-01  7.518e-01  -1.313
## `native.countryDominican-Republic`         -2.992e+00  1.241e+00  -2.411
## native.countryEcuador                      -1.984e+00  1.090e+00  -1.821
## `native.countryEl-Salvador`                -1.993e+00  8.581e-01  -2.323
## native.countryEngland                      -1.136e+00  7.519e-01  -1.511
## native.countryFrance                       -7.782e-01  8.396e-01  -0.927
## native.countryGermany                      -9.965e-01  7.178e-01  -1.388
## native.countryGreece                       -2.397e+00  9.061e-01  -2.645
## native.countryGuatemala                    -1.481e+00  1.012e+00  -1.464
## native.countryHaiti                        -1.164e+00  9.685e-01  -1.202
## `native.countryHoland-Netherlands`                 NA         NA      NA
## native.countryHonduras                     -2.492e+00  2.736e+00  -0.911
## native.countryHong                         -2.609e+00  1.259e+00  -2.072
## native.countryHungary                      -1.857e+00  1.128e+00  -1.645
## native.countryIndia                        -1.760e+00  7.080e-01  -2.485
## native.countryIran                         -1.215e+00  8.176e-01  -1.486
## native.countryIreland                      -1.646e+00  1.090e+00  -1.510
## native.countryItaly                        -5.996e-01  7.514e-01  -0.798
## native.countryJamaica                      -1.447e+00  8.330e-01  -1.737
## native.countryJapan                        -1.254e+00  7.766e-01  -1.615
## native.countryLaos                         -9.284e-01  1.173e+00  -0.792
## native.countryMexico                       -2.022e+00  7.131e-01  -2.835
## native.countryNicaragua                    -1.977e+00  1.049e+00  -1.884
## `native.countryOutlying-US(Guam-USVI-etc)` -1.462e+01  3.730e+02  -0.039
## native.countryPeru                         -1.450e+01  2.503e+02  -0.058
## native.countryPhilippines                  -9.121e-01  6.844e-01  -1.333
## native.countryPoland                       -1.193e+00  7.814e-01  -1.527
## native.countryPortugal                     -8.733e-01  9.628e-01  -0.907
## `native.countryPuerto-Rico`                -1.751e+00  7.839e-01  -2.234
## native.countryScotland                     -1.707e+00  1.325e+00  -1.288
## native.countrySouth                        -2.461e+00  7.817e-01  -3.148
## native.countryTaiwan                       -1.465e+00  8.287e-01  -1.768
## native.countryThailand                     -1.974e+00  1.042e+00  -1.895
## `native.countryTrinadad&Tobago`            -1.506e+00  1.123e+00  -1.341
## `native.countryUnited-States`              -1.171e+00  6.642e-01  -1.763
## native.countryVietnam                      -1.918e+00  8.849e-01  -2.167
```

```
## native.countryYugoslavia                  -7.121e-01  1.006e+00  -0.708
##                                           Pr(>|z|)
## (Intercept)                               7.99e-13 ***
## age                                        < 2e-16 ***
## `workclassLocal-gov`                      5.71e-08 ***
## workclassPrivate                          7.81e-06 ***
## `workclassSelf-emp-inc`                   0.033911 *
## `workclassSelf-emp-not-inc`               1.57e-15 ***
## `workclassState-gov`                      3.10e-09 ***
## `workclassWithout-pay`                    0.967222
## fnlwgt                                    0.000632 ***
## education11th                             0.432702
## education12th                             0.253307
## `education1st-4th`                        0.420263
## `education5th-6th`                        0.417783
## `education7th-8th`                        0.017508 *
## education9th                              0.493051
## `educationAssoc-acdm`                     3.97e-10 ***
## `educationAssoc-voc`                      6.85e-11 ***
## educationBachelors                         < 2e-16 ***
## educationDoctorate                         < 2e-16 ***
## `educationHS-grad`                        6.64e-06 ***
## educationMasters                           < 2e-16 ***
## educationPreschool                        0.905815
## `educationProf-school`                     < 2e-16 ***
## `educationSome-college`                   4.26e-10 ***
## education.num                                   NA
## `marital.statusMarried-AF-spouse`         3.27e-06 ***
## `marital.statusMarried-civ-spouse`        6.82e-11 ***
## `marital.statusMarried-spouse-absent`     0.919428
## `marital.statusNever-married`             1.47e-07 ***
## marital.statusSeparated                   0.554061
## marital.statusWidowed                     0.300111
## `occupationArmed-Forces`                  0.466845
## `occupationCraft-repair`                  0.601993
## `occupationExec-managerial`                < 2e-16 ***
## `occupationFarming-fishing`               1.11e-08 ***
## `occupationHandlers-cleaners`             9.02e-06 ***
## `occupationMachine-op-inspct`             0.008238 **
## `occupationOther-service`                 1.61e-09 ***
## `occupationPriv-house-serv`               0.035882 *
## `occupationProf-specialty`                1.34e-08 ***
## `occupationProtective-serv`               4.16e-05 ***
## occupationSales                           0.011451 *
## `occupationTech-support`                  5.71e-07 ***
## `occupationTransport-moving`              0.501098
## `relationshipNot-in-family`               0.259619
## `relationshipOther-relative`              0.052788 .
## `relationshipOwn-child`                   0.008096 **
## relationshipUnmarried                     0.426088
## relationshipWife                           < 2e-16 ***
## `raceAsian-Pac-Islander`                  0.070538 .
## raceBlack                                 0.404448
## raceOther                                 0.995745
```

```
## raceWhite                                    0.135040
## sexMale                                       < 2e-16 ***
## capital.gain                                  < 2e-16 ***
## capital.loss                                  < 2e-16 ***
## hours.per.week                                < 2e-16 ***
## native.countryCanada                         0.154145
## native.countryChina                          0.017085 *
## native.countryColumbia                       0.002582 **
## native.countryCuba                           0.189306
## `native.countryDominican-Republic`           0.015929 *
## native.countryEcuador                        0.068671 .
## `native.countryEl-Salvador`                  0.020171 *
## native.countryEngland                        0.130713
## native.countryFrance                         0.354006
## native.countryGermany                        0.165019
## native.countryGreece                         0.008161 **
## native.countryGuatemala                      0.143274
## native.countryHaiti                          0.229526
## `native.countryHoland-Netherlands`                 NA
## native.countryHonduras                       0.362426
## native.countryHong                           0.038241 *
## native.countryHungary                        0.099873 .
## native.countryIndia                          0.012943 *
## native.countryIran                           0.137239
## native.countryIreland                        0.130947
## native.countryItaly                          0.424922
## native.countryJamaica                        0.082381 .
## native.countryJapan                          0.106292
## native.countryLaos                           0.428477
## native.countryMexico                         0.004576 **
## native.countryNicaragua                      0.059544 .
## `native.countryOutlying-US(Guam-USVI-etc)` 0.968741
## native.countryPeru                           0.953805
## native.countryPhilippines                    0.182588
## native.countryPoland                         0.126847
## native.countryPortugal                       0.364392
## `native.countryPuerto-Rico`                  0.025504 *
## native.countryScotland                       0.197707
## native.countrySouth                          0.001646 **
## native.countryTaiwan                         0.077059 .
## native.countryThailand                       0.058142 .
## `native.countryTrinadad&Tobago`              0.180072
## `native.countryUnited-States`                0.077925 .
## native.countryVietnam                        0.030213 *
## native.countryYugoslavia                     0.478932
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27079  on 24128  degrees of freedom
## Residual deviance: 15598  on 24034  degrees of freedom
## AIC: 15788
##
```

```
## Number of Fisher Scoring iterations: 14
```