Adult Census Income Project

Pim Kempe

March 2020

Contents

1 Overview

This project is the second assignment of the online edX course *HarvardX: PH125.9x - Data Science: Capstone*, which is the final part of the program *HarvardX Data Science Professional Certificate*. The goal of this assignment is to apply the knowledge gained during this program and use a publicly available data set to apply machine learning techniques.

1.1 Introduction

In this assignment the Adult Census Income data set is used. This a publicly available data set on Kaggle created Ronny Kohavi and Barry Becker. The data is extracted from the 1994 Census bureau database and contains demographic information such as age, sex and education. The task of this project is to predict whether an individual's income exceeds 50.000 US Dollar per year. Hence this is a binary classification problem.

The data set can be downloaded here and is reasonably clean and ready to use. More information can be found on Kaggle.

This project is carried out with *R* and *RStudio*. *R* is an environment and programming language focused on statistical analysis. *RStudio* is an environment for creating scripts and visualizations using the *R* language.

2 Methods & Analysis

2.1 Data Introduction

The Adult Census Income data set contains 32561 observations and 15variables. The variables are both numerical and categorical, and are outlined below:

- Response variable:
  - **income**: Categorical variable that contains yearly income of the respondent ("<=$50K" or ">50K").
- Features:
  - **age**: Numerical variable that contains the age of the respondent.
  - **workclass**: Categorical variable that contains the type of employer of the respondent:
    - *?*: Unknown
    - *Federal-gov*: Federal Government
    - *Local-gov*: Local Government
    - *Never-worked*: Never Worked
    - *Private*: Private Sector
    - *Self-emp-inc*: Self Employment (Corporate Entities)
    - *Self-emp-not-inc*: Self Employment (Other Legal Entities)
    - *State-gov*: State Government
    - *Without-pay*: Unemployed
  - **fnlwgt**: Numerical variable that contains the number of respondents that each row of the data set represents.
  - **education**: Categorical variable that represents the level of education of the respondent (*Doctorate*, *Prof-school*, *Masters*, *Bachelors*, *Assoc-acdm*, *Assoc-voc*, *Some-college*, *HS-grad*, *12th*, *11th*, *10th*, *9th*, *7th-8th*, *5th-6th*, *1st-4th*, *Preschool*).
  - **education.num**: Numerical variable that represents the *education* variable.
  - **marital.status**: The marital status of the respondent:
    - *Divorced*: Divorced
    - *Married-AF-spouse*: Married (Armed Forces spouse)
    - *Married-civ-spouse*: Married (civil spouse)
    - *Married-spouse-absent*: Married (living without spouse)
    - *Never-married*: Never married
    - *Separated*: Separated
    - *Widowed*: Widowed
  - **occupation**: Categorical variable that represents the type of employment of the respondent (*?*, *Adm-clerical*, *Armed-Forces*, *Craft-repair*, *Exec-managerial*, *Farming-fishing*, *Handlers-cleaners*, *Machine-op-inspct*, *Other-service*, *Priv-house-serv*, *Prof-specialty*, *Protective-serv*, *Sales*, *Tech-support*, *Transport-moving*).
  - **relationship**: Categorical variable that represents the position in the family of the respondent (*Husband*, *Not-in-family*, *Other-relative*, *Own-child*, *Unmarried*, *Wife*).
  - **race**: Categorical variable that represents the race of the respondent (*Amer-Indian-Eskimo*, *Asian-Pac-Islander*, *Black*, *Other*, *White*).
  - **sex**: Categorical variable that represent the sex of the respondent (*Female*, *Male*).
  - **capital.gain**: Numerical variable that represents the income gained by the respondent from sources other than salary/wages.
  - **capital.loss**: Numerical variable that represents the income lost by the respondent from sources other than salary/wages.
  - **hours.per.week**: Numerical variable that represents the hours worked per week by the respondent.
  - **native.country**: Categorical variable that represents the native country of the respondent.

2.2 Data Preparation

The figure below shows a snippet of the raw Adult Census Income data set.

age	workclass	fnlwgt	education	education.num	marital.status	occupation
90	?	77053	HS-grad	9	Widowed	?
82	Private	132870	HS-grad	9	Widowed	Exec-managerial
66	?	186061	Some-college	10	Widowed	?
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct
74	State-gov	88638	Doctorate	16	Never-married	Prof-specialty
68	Federal-gov	422013	HS-grad	9	Divorced	Prof-specialty
41	Private	70037	Some-college	10	Never-married	Craft-repair
45	Private	172274	Doctorate	16	Divorced	Prof-specialty
38	Self-emp-not-inc	164526	Prof-school	15	Never-married	Prof-specialty

relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
Not-in-family	White	Female	0	4356	40	United-States	<=50K
Not-in-family	White	Female	0	4356	18	United-States	<=50K
Unmarried	Black	Female	0	4356	40	United-States	<=50K
Unmarried	White	Female	0	3900	40	United-States	<=50K
Other-relative	White	Female	0	3683	20	United-States	>50K
Not-in-family	White	Female	0	3683	40	United-States	<=50K
Unmarried	White	Male	0	3004	60	?	>50K
Unmarried	Black	Female	0	3004	35	United-States	>50K
Not-in-family	White	Male	0	2824	45	United-States	>50K

As can be observed, missing/unknown values are represented by a '?'. Since *R* doesn't recognize a '?' as a missing/unknown value, the '?'s are replaced by 'N/A's.

The variables *fnlwgt* and *education* are removed from the data set. The variable *fnlwgt* represents the weight of a specific combination of demographic characteristics in the data. Since this variable is merely a descriptive statistic of the census data, it can't/shouldn't be used to predict whether an individual earns more than 50.000 US Dollar per year. The variable *education* is removed because it's just a textual representation of the variable *education.num*. Remaining both *education* and *education.num* in the data set is useless.

The response variable *income* will be changed such that "<=50K" = 0 and ">50K" = 1.

Finally, the data set is separated into a training set (70%) and a validation set (30%). All data analysis and modelling is done on the training set since the validation set is ought to be considered as unknown. The validation set is only used to assess the performance of the final models.

2.3 Exploratory Data Analysis

As mentioned above the data set is separated into a training set and test set. All Exploratory Data Analysis in this section is carried out using the training set.

The training set contains 22792 observations of which 24% has an income that exceeds 50.000 US Dollar per year. In the remaining part of this section each variable will be investigated separately.

2.3.1 Age

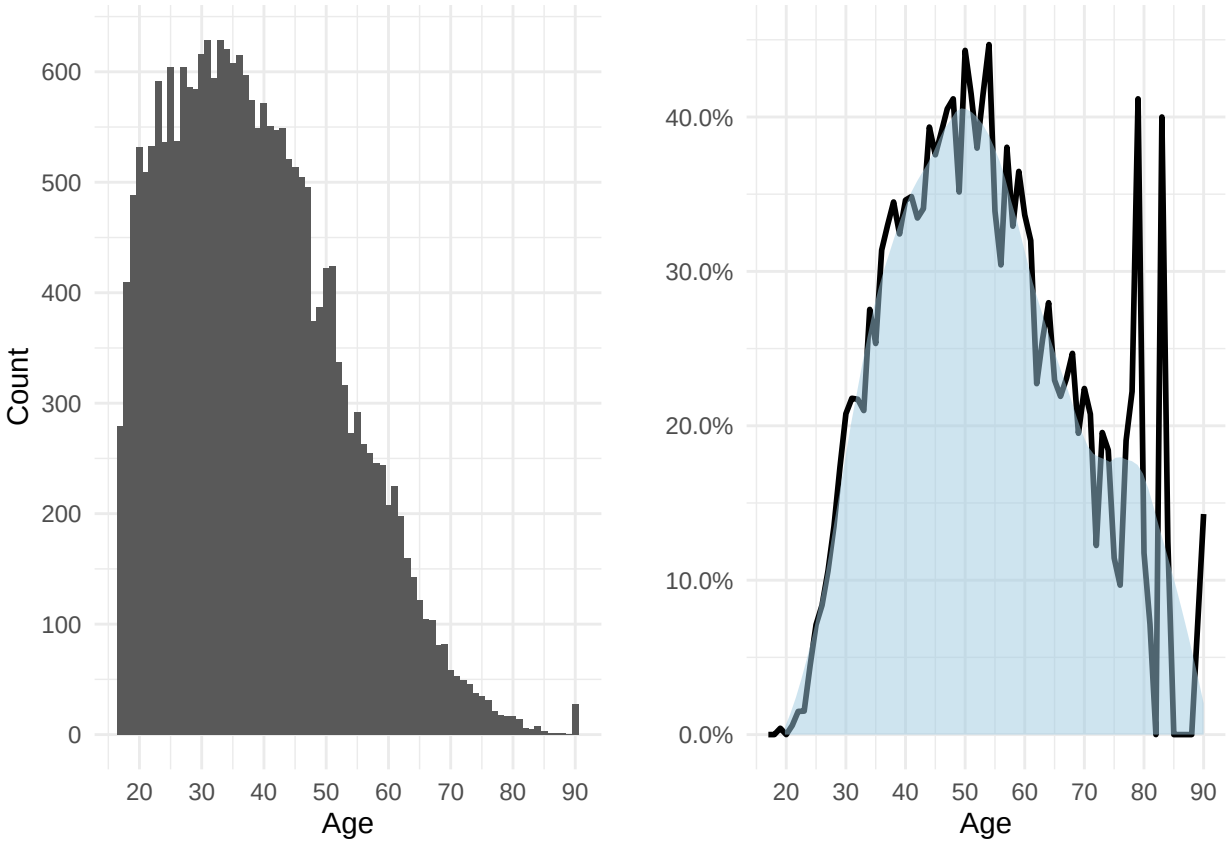

Figure 1: Age

Figure 1 shows the distribution of the respondents age (left) and the percentage of respondents that earns more than 50.000 US Dollar per year by age (right). The average age of the respondent is 39 years. Respondent between 35 and 60 years are more likely to earn more than 50.00 US Dollar per year.

2.3.2 Workclass

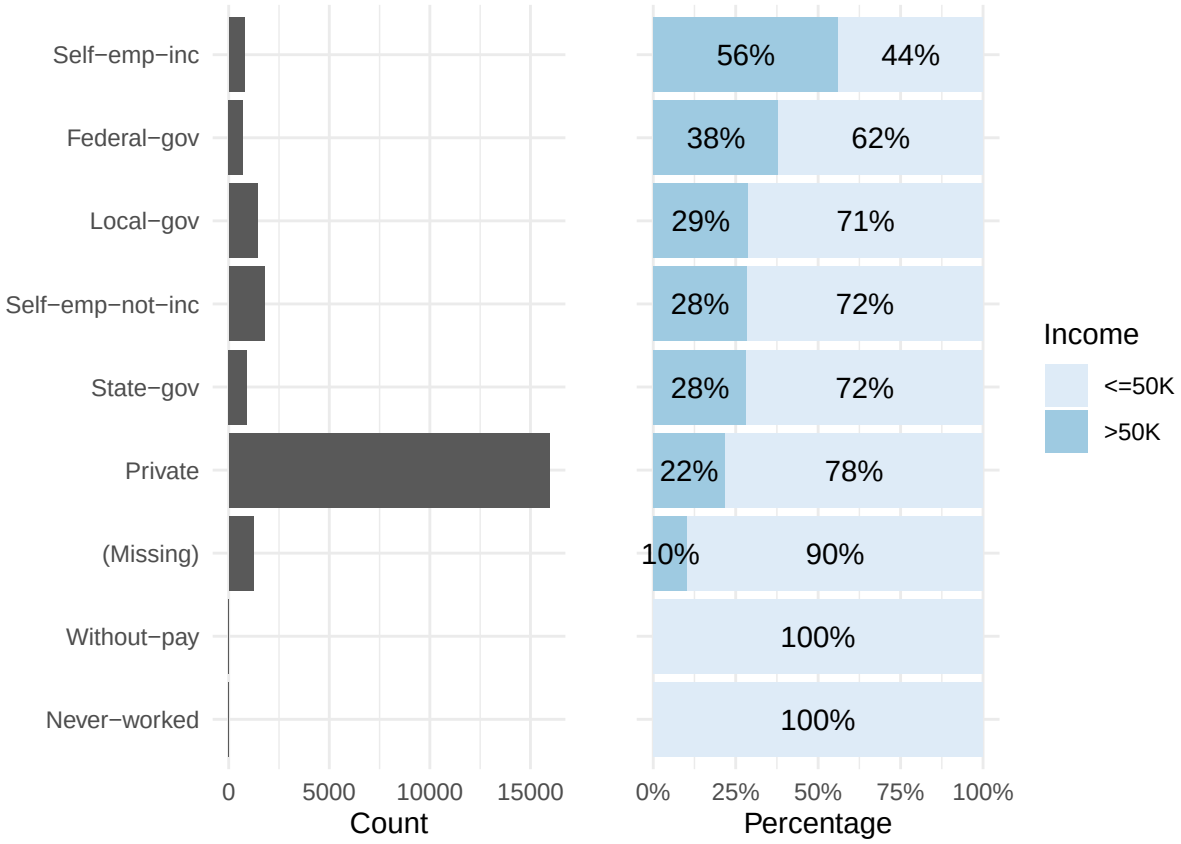

Figure 2: Workclass

Figure 2 shows that most respondents work for an employer in the private sector. Respondents that work for the government or are self-employed are more likely to earn more than 50.000 US Dollar per year.

2.3.3 Education

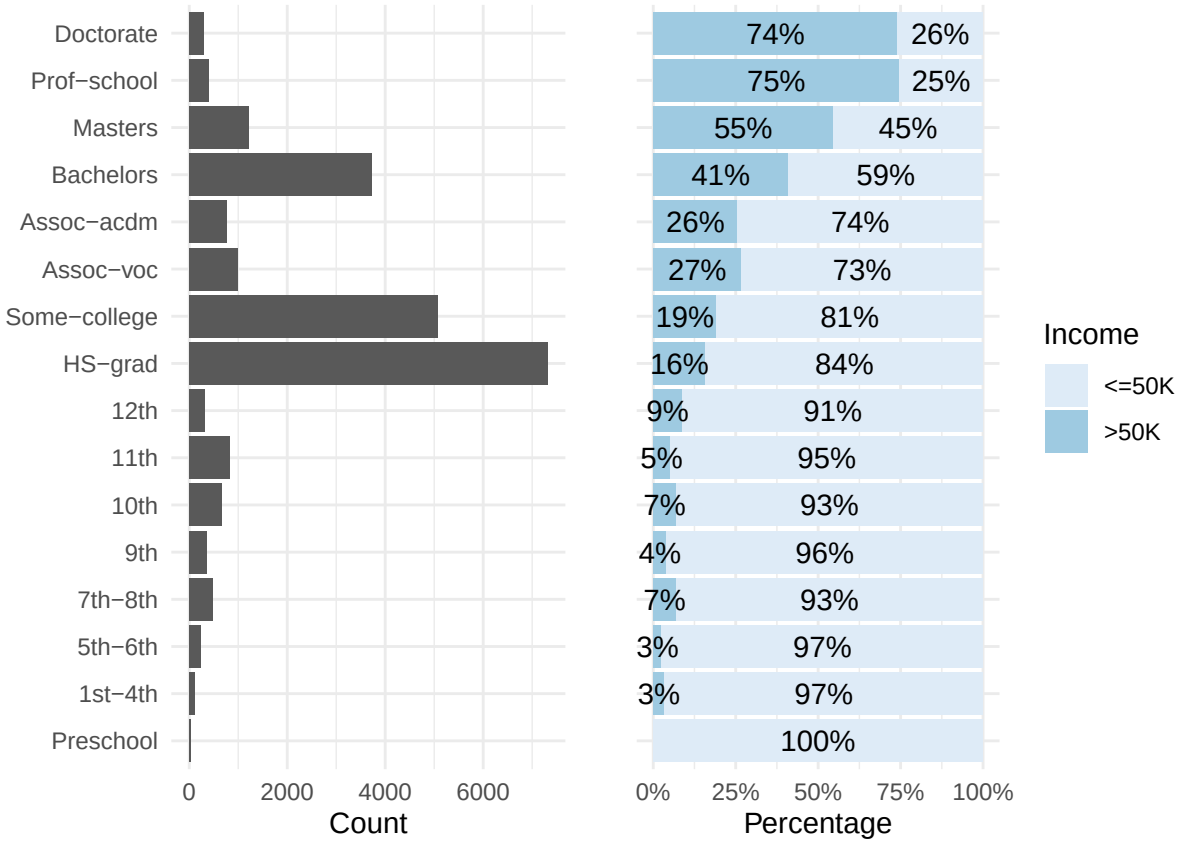

Figure 3: Education

Figure 3 shows that most of the respondents finished High School, did some college or have a Bachelors degree. Respondents that finished an educational level higher than college are more likely to earn more than 50.000 US Dollar per year than average (24%).

2.3.4 Marital Status

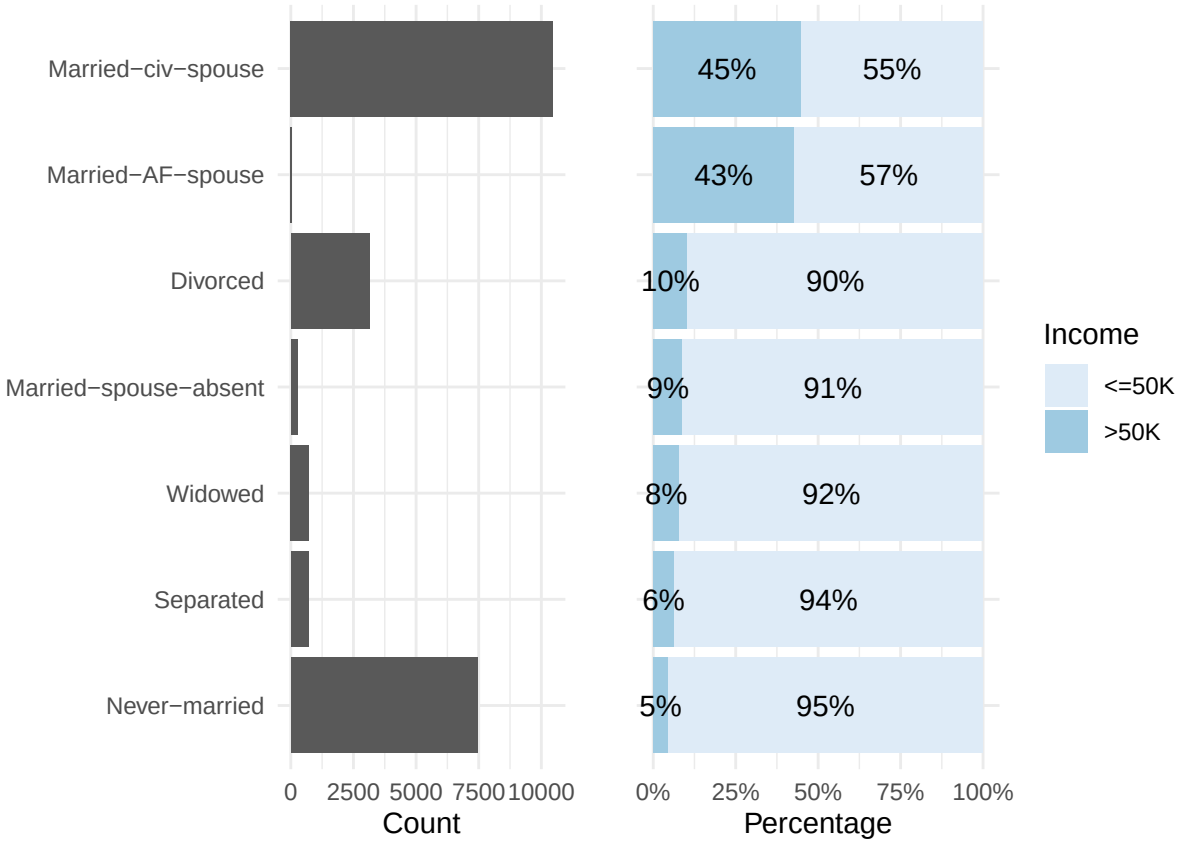

Figure 4: Marital Status

Figure 4 shows that most of the respondents are currently married or never have been married. Married respondents are much more likely to earn more than 50.000 US Dollar per year than non-married respondents.

2.3.5 Occupation

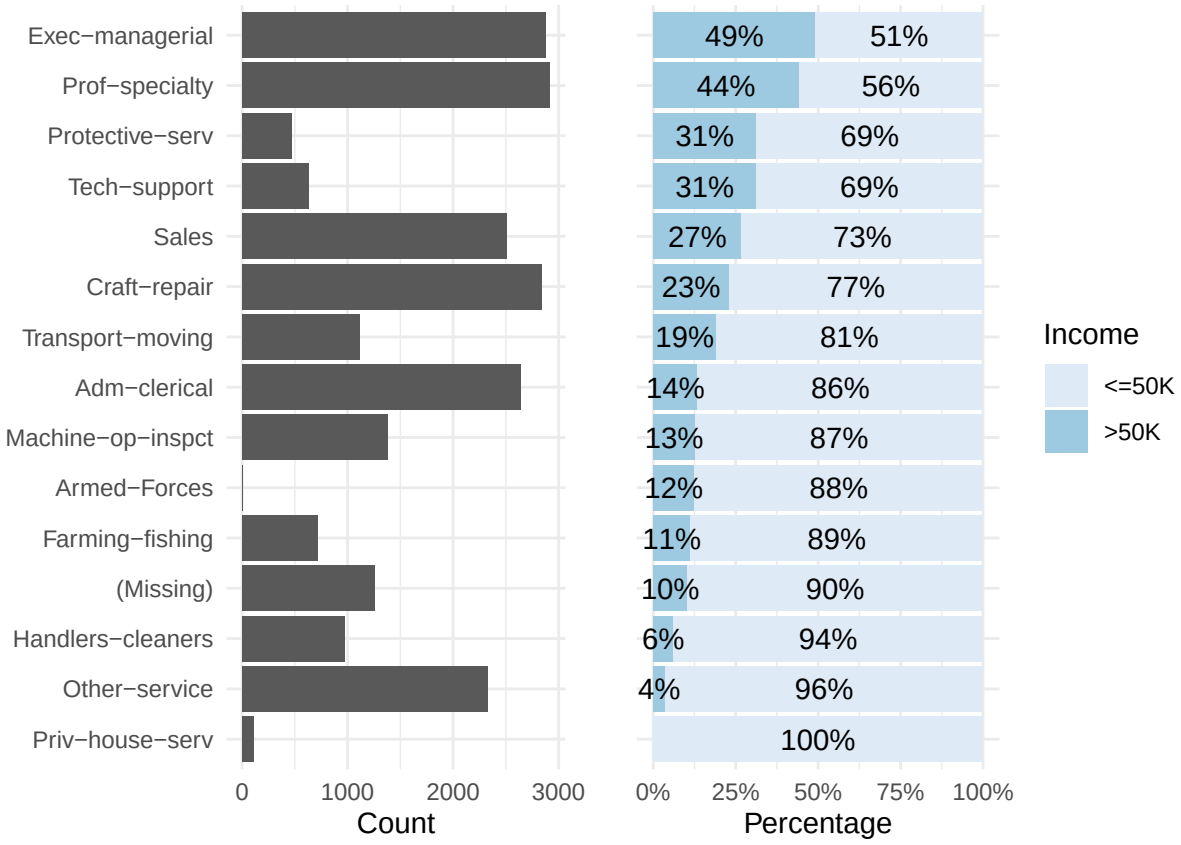

Figure 5: Occupation

Figure 5 shows the distribution of the respondent's occupation. It can be observed that respondents with a position the the executive management or a professional specialization have a higher than average probability to gain more than 50.000 US Dollar per year.

2.3.6 Relationship

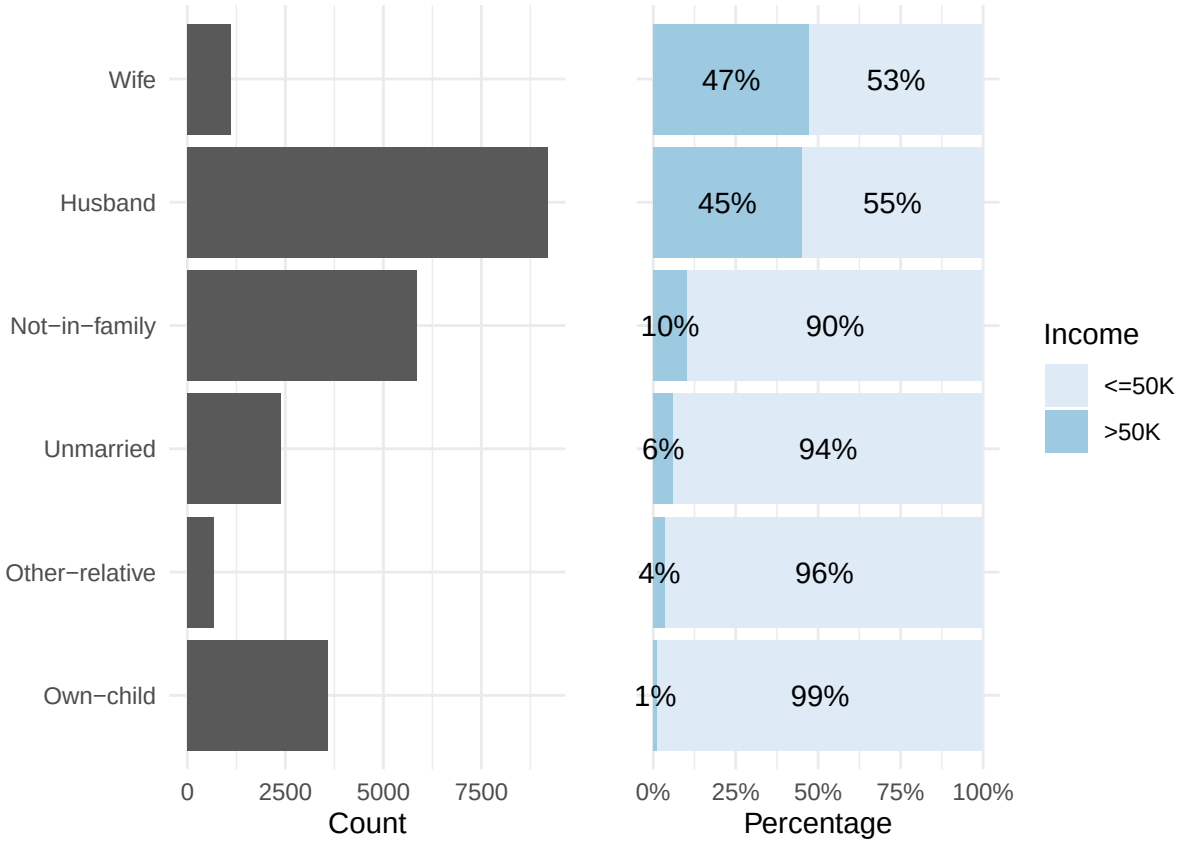

Figure 6: Relationship

Figure 6 shows the distribution of the respondent's relationship. In concordance with Figure 4, married respondents (Wife or Husband) have a higher probability than average to earn more than 50.000 US Dollar per year.

2.3.7 Race

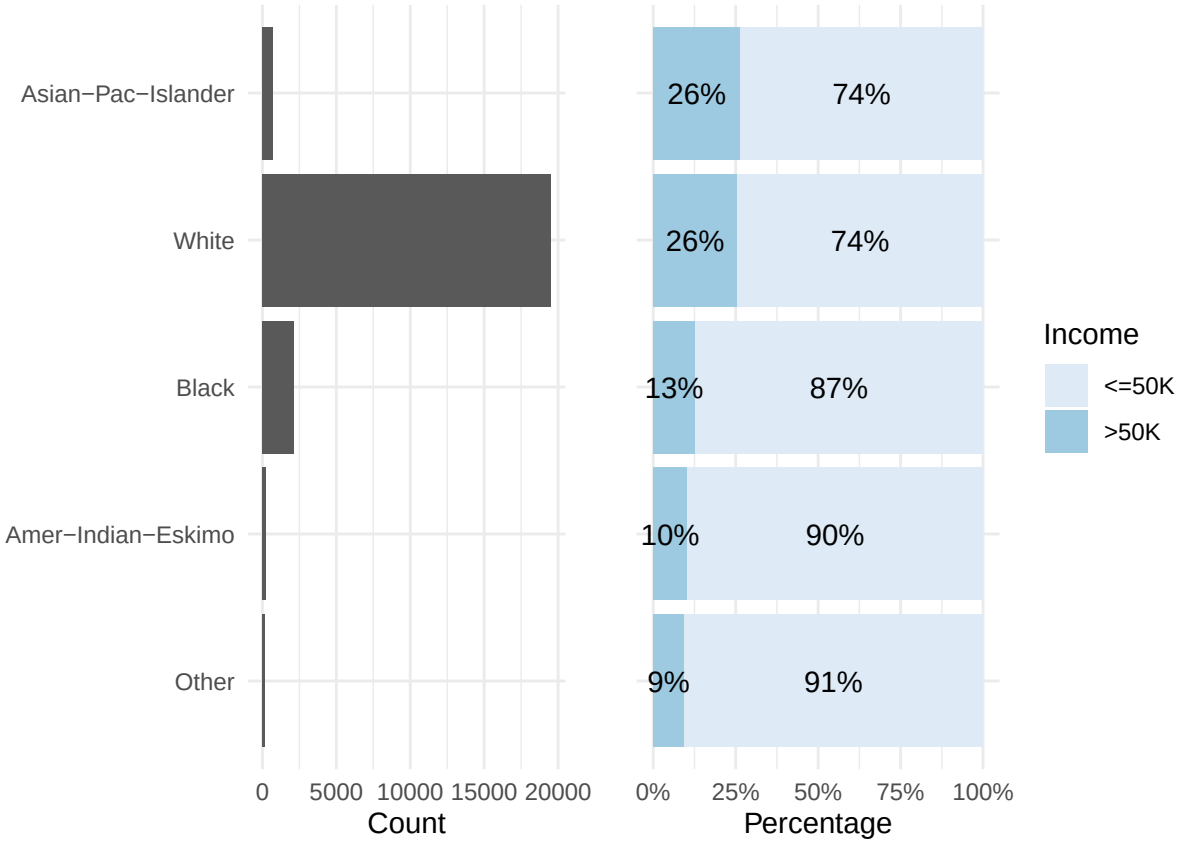

Figure 7: Race

Figure 7 shows the distribution of the respondent's race. Since most respondents are white this variable might not be very useful whether a respondent earns more than 50.000 US Dollar per year.

2.3.8 Sex

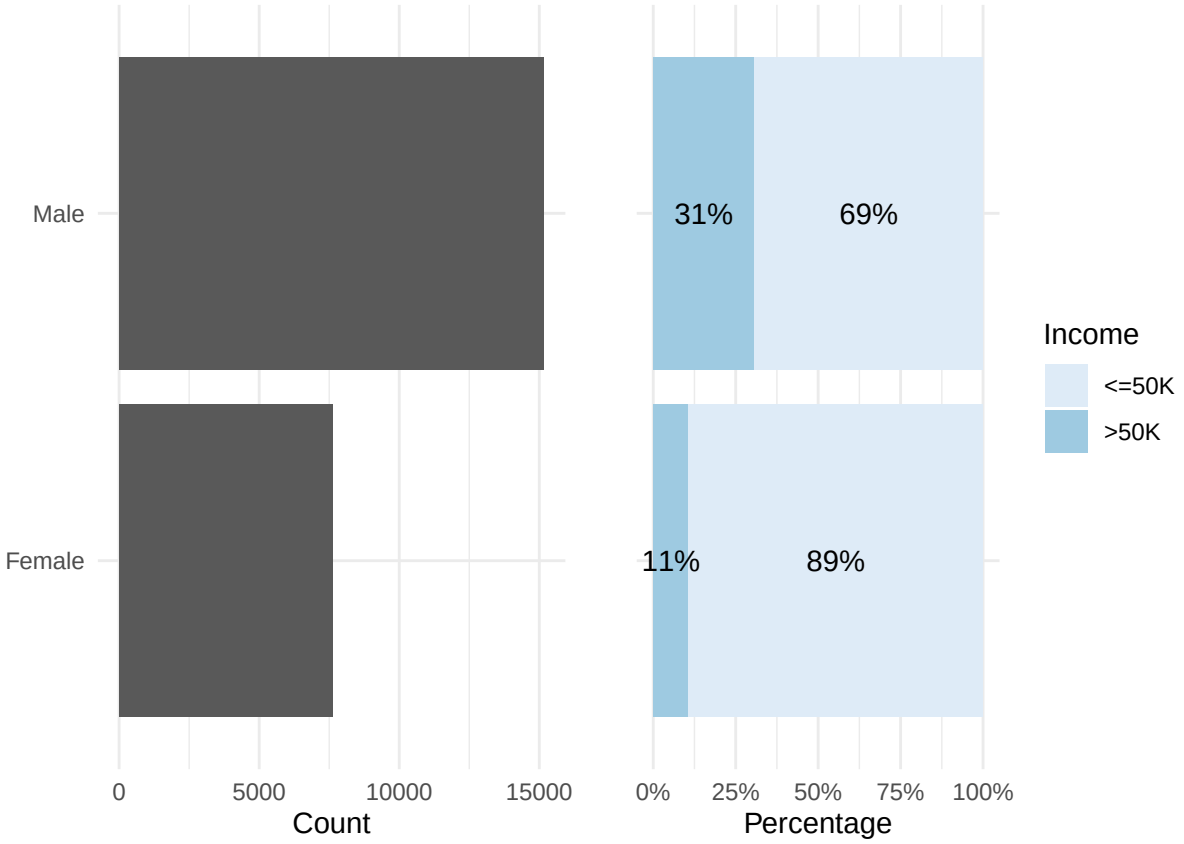

Figure 8: Sex

Figure 8 shows that approximately two-third of the respondents is male. It also shows that males are more likely to earn more then 50.000 US Dollar per year than females.

2.3.9 Capital Gain/Loss

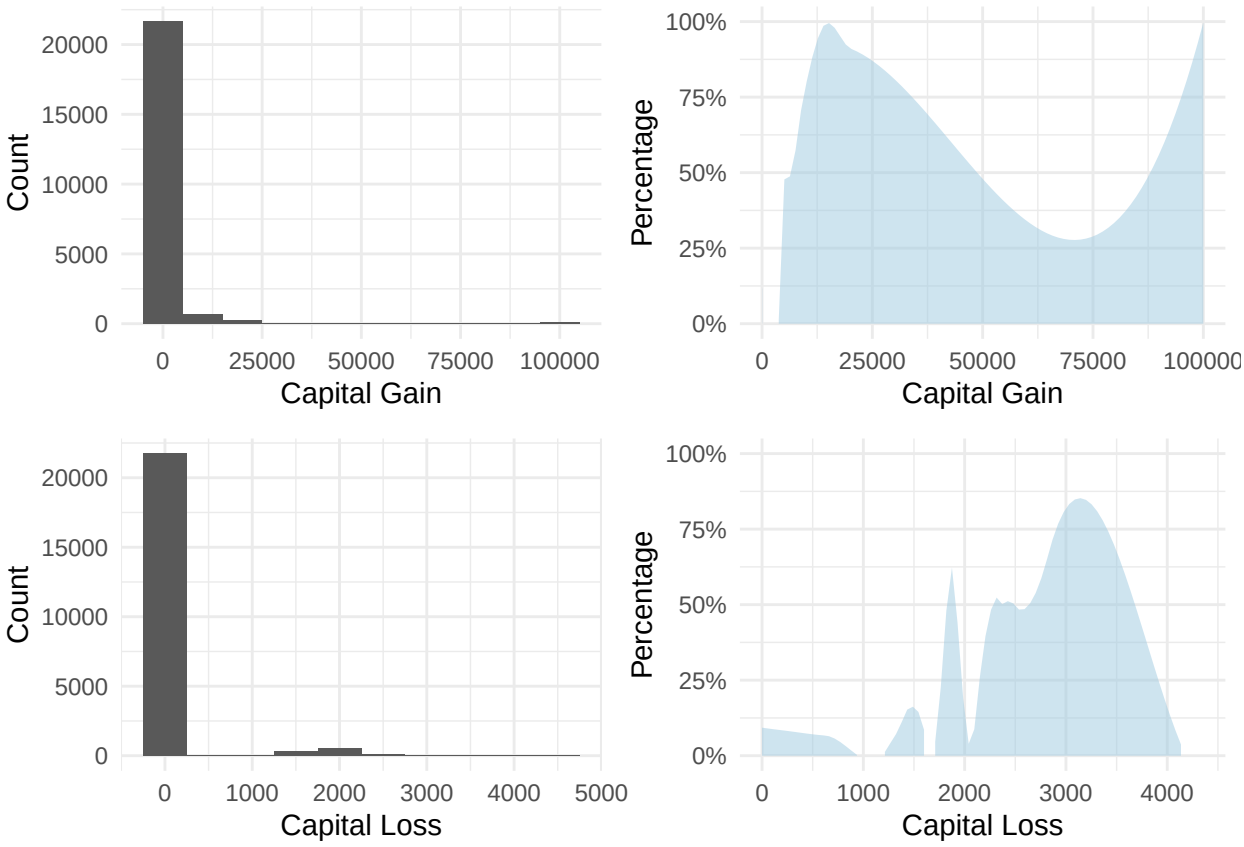

Figure 9: Captail Gain/Loss

Figure 9 shows the income that the respondent gained/lost from sources than salary/wages. It's clear that most respondents don't have other income sources than salary/wages. Those who have other income sources are likely to gain more than 50.000 US Dollar per year.

2.3.10 Hours per Week

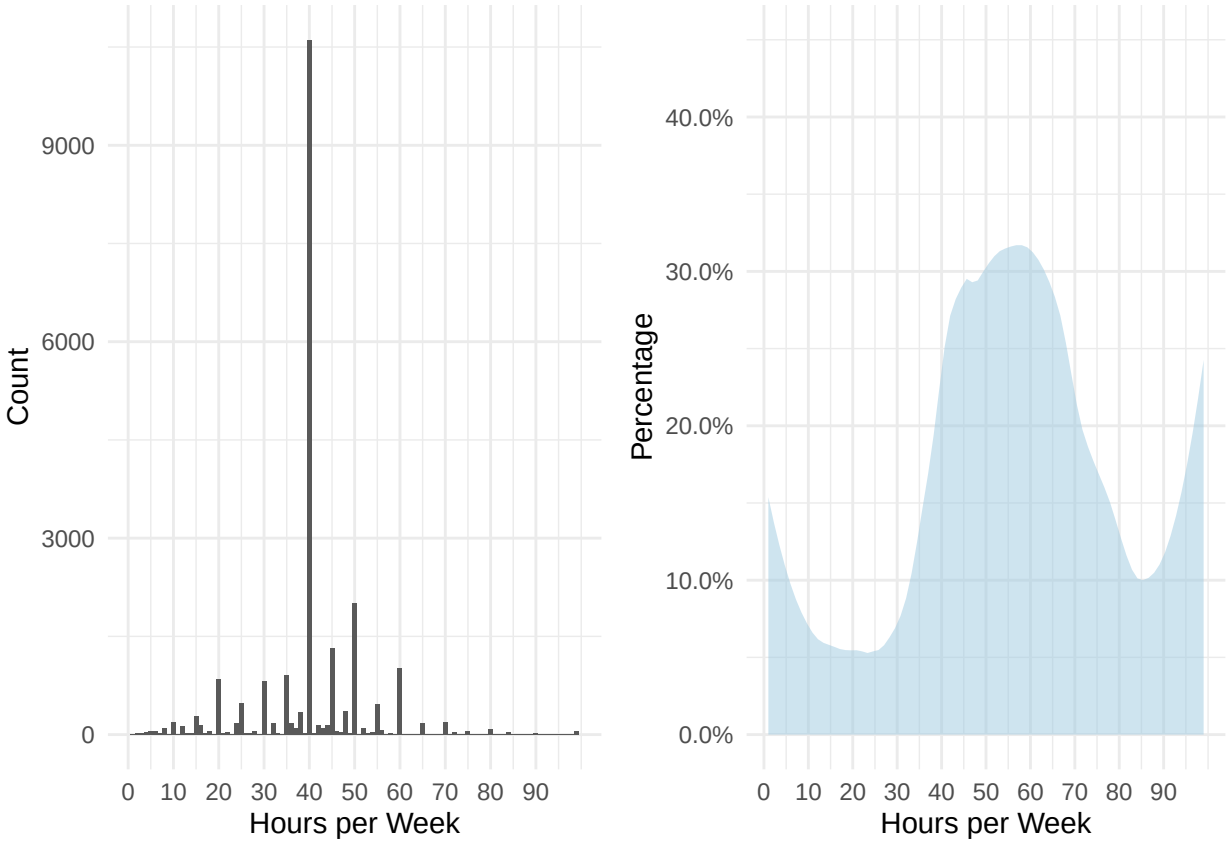

Figure 10: Hours per Week

Figure 10 shows that most respondents work 40 hours per week. Respondents that earn more than 50.000 US Dollar per year work mostly likely more than 40 hours per week and less than 70 hours per week.

2.3.11 Native Country

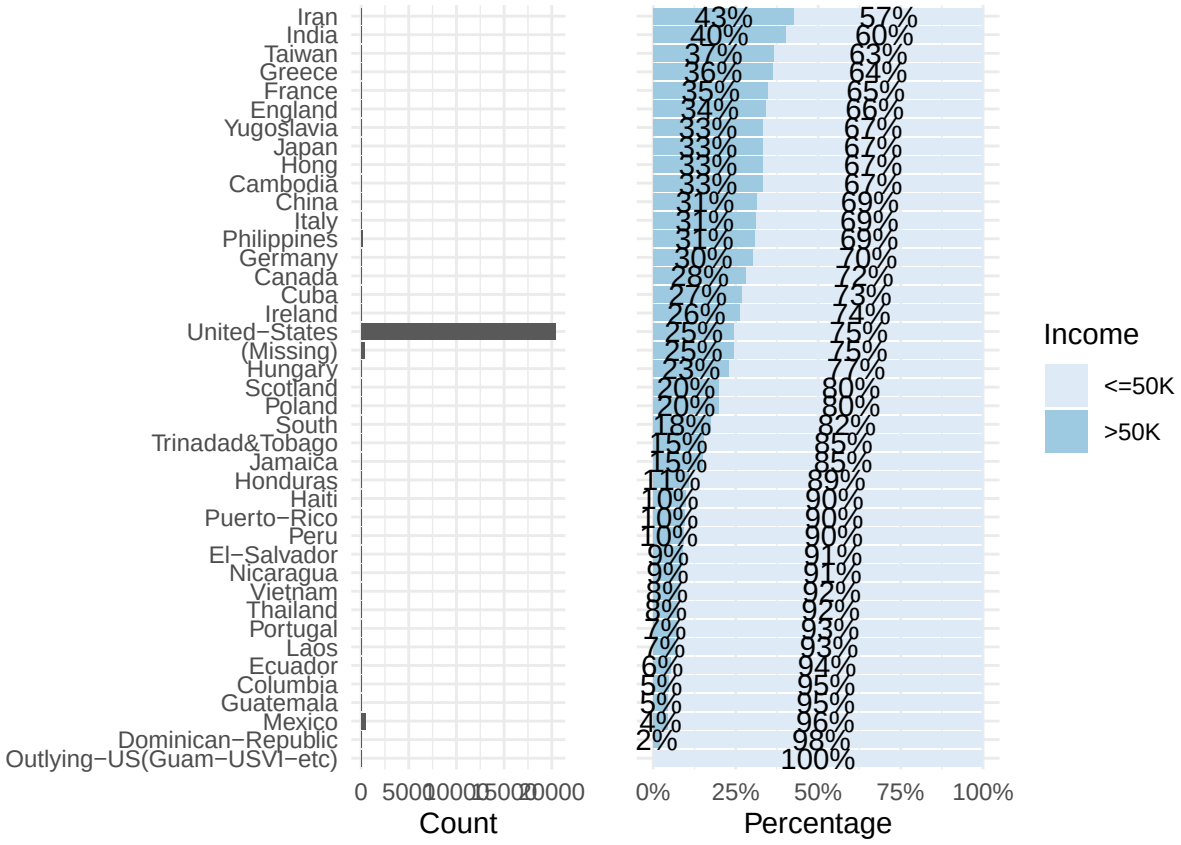

Figure 11: Native Country

Figure 11 shows that almost all respondents are from the United States. This variable will probably not useful in predicting whether a respondent will earn more than 50.000 US Dollar per year.

2.3.12 Exploratory Data Analysis - Conclusions

As a result of the Exploratory Data Analysis insights are gained and four new columns are created:

- **married**: Categorical variable that indicates whether a respondent is currently married (*Married* = 1, *Not-Married* = 0).
- **other.income**: Categorical variable that indicates whether a respondent has another source of income besides salary/wages. If a respondent has a Capital Gain > 0 or a Capital Loss > 0 it's assumed (s)he has another source of income. (*Other Income* = 1, *No Other Income* = 0).
- **hours.per.week.2**: Categorical variable that indicates whether a respondent work more than 40 hours per week ($>$ *40hrs p/w* = 1, $<=$ *40hrs p/w* = 0).
- **workclass.2**: Categorical variable based on the variable *workclass*. This new variable contains the categories *Private*, *Government*, *self Employed* and *Other*.

2.4 Modeling Methods

The project will use two methods to predict whether a respondent gain more than 50.000 US Dollar per year.

2.4.1 k-Nearest Neighbors

The first method that will be used is K-Nearest Neighbors. This is an intuitive and fast method for classification problems.

The K-Nearest Neighbors method classifies new observations based on its similarity to known observations. The similarity is based on the distance between the variables of a new observation and the variables of the k closest known observations.

Since the distance can only be measured for numeric variables the following variables are included in the model:

- **age**
- **education.num**
- **capital.gain**
- **capital.loss**
- **hours.per.week**

The variables are pre-processed before the modelling takes place because the different ranges of the variables can distort the classification process. For example, the variable *education.num* ranges between 1 and 16, while *capital.gain* has a range of 0 to 99999. All numeric variables are centered and scales, i.e. the observations value is extracted by it's mean and divided by it's standard deviation (of all observations).

There exist different definitions to measure the distance between observation; Euclidean distance, Manhattan distance, Minkwoski distance, etc. In this project the default method (Euclidean) is used to measure the distance. The *caret* package is used to train the model.

10-Fold Cross Validation is used to train the model. This is repeated three times. In total 13 models with a different number of neighbors were trained. Figure 12 shows the accuracy of the different models using 10-fold Cross Validation.

The trained model with 23 neighbors has the highest accuracy. Although the accuracy with 23 neighbors looks a bit 'extreme' compared to the nearby neighbors and the ascend in the accuracy seems to stop near 30 neighbors, this model (K=23) is still used to measure the performance of the validation set.

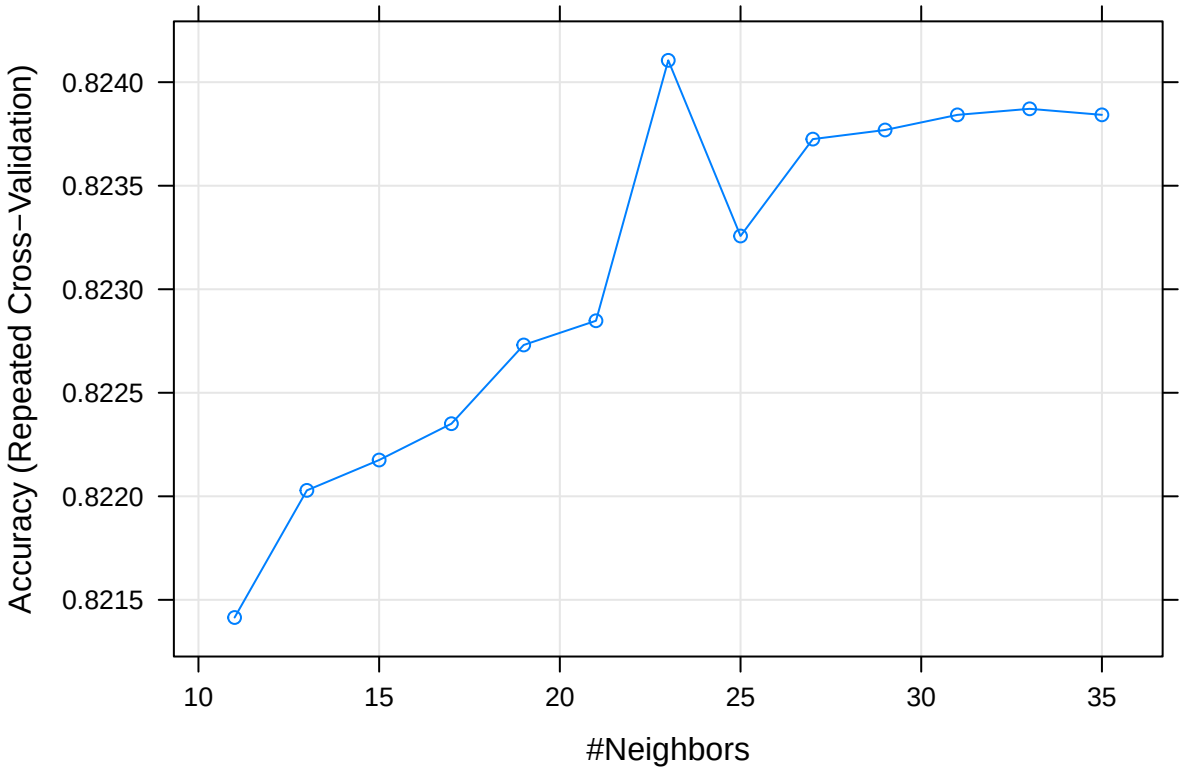

Figure 12: Training a k-NN model

2.4.2 Random Forest

The second method that will be used is a Random Forest. A Random Forest is an ensemble classification method were different uncorrelated decision trees predict by committee the class of a new observation.

Random Forests can handle both numeric as categorical variables. However, training a Random Forest can be time consuming (on a normal consumer laptop). Based on the results of the Exploratory Data Analysis section a selection has been made of variable that will be included in the model and reduce training time. The variables that are included based on how well they separate the two classes (<=50K p/y or >50K p/y). Figures 1 - 11 helped to select the following variables:

- **age**
- **education.num**
- **marital.status**
- **relationship**
- **capital.gain**
- **sex**

The *randomForest* package in *R* provides the user a lot of options for parameter tuning, such as the number of decision trees used in the forest, the maximum depth of each tree, etc. However, since training a Random Forest is time quite time consuming all default parameters were used.

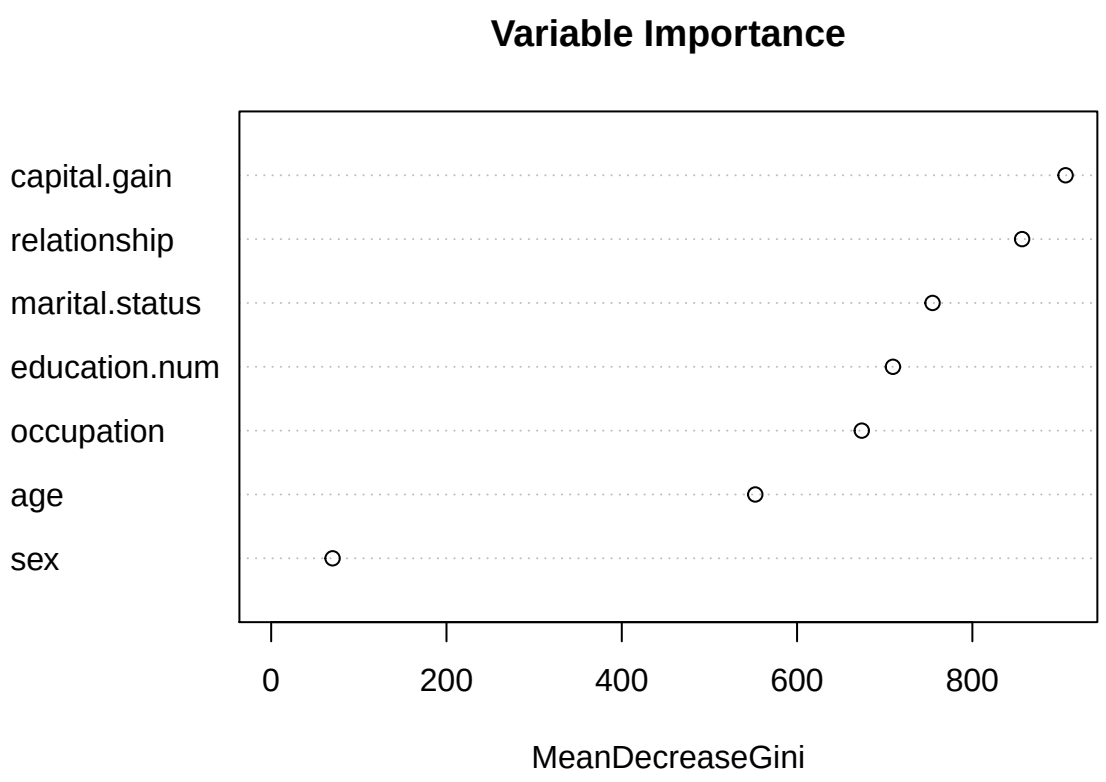

Figure 13: Training a Random Forest

Figure 13 the importance of each variable in the Random Forest. It turns out that the *relationship* and *capital.gain* variables are the most important to classify new observations. This is in line with Figures 6 and 9.

3 Results

In this section the performance of the k-Nearest Neighbors model and the Random Forest is compared.

The performance of the selected k-Nearest Neighbors model (K=23) on the validation set is shown below. The model has an accuracy of 0.8219. In the confusion matrix Class 1 represents the respondents that earn more than 50.000 US Dollar per year.

In this project we're interested in classifying correctly the respondents that earn more than 50.000 US Dollar per year (Class 1). It can be observed that only 1053 of the 2353 respondents of Class 1 are classified correctly, i.e. a specificity of 0.4475.

The performance of the Random Forest on the validation set is shown below. Again, Class 1 are the respondents that earn more than 50.000 US Dollar per year. The Random Forest has a higher accuracy than the k-Nearest Neighbors model, 0.8536 vs. 0.8219 respectively. With a specificity of 0.6022 the Random Forest has a higher specificity than the k-Nearest Neighbors model.

Comparing the 95% confidence intervals of the accuracy of the k-Nearest Neighbors model (0.8142 - 0.8294) and the Random Forest (0.8465 - 0.8606) it can be concluded that the Random Forest performs better.

```
## [1] "Performance k-Nearest Neighbors"

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 6976 1302
##          1  440 1051
##
##                Accuracy : 0.8217
##                  95% CI : (0.8139, 0.8292)
##     No Information Rate : 0.7591
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4427
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9407
##             Specificity : 0.4467
##          Pos Pred Value : 0.8427
##          Neg Pred Value : 0.7049
##              Prevalence : 0.7591
##          Detection Rate : 0.7141
##    Detection Prevalence : 0.8474
##       Balanced Accuracy : 0.6937
##
##        'Positive' Class : 0
##

## [1] "Performance Random Forest"

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 6922  932
##          1  494 1421
##
##                Accuracy : 0.854
##                  95% CI : (0.8469, 0.861)
##     No Information Rate : 0.7591
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5738
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9334
##             Specificity : 0.6039
##          Pos Pred Value : 0.8813
##          Neg Pred Value : 0.7420
##              Prevalence : 0.7591
```

```
##          Detection Rate : 0.7086
##    Detection Prevalence : 0.8040
##       Balanced Accuracy : 0.7686
##
##        'Positive' Class : 0
##
```

4 Discussion

In this assignment the publicly available Adult Census Income data was used to predict whether a respondent earns more than 50.000 US Dollar per year.

First an Exploratory Data Analysis was carried out. Then two machine learning models were build; a k-Nearest Neighbors model and a Random Forest. The Random Forest model performed better (higher accuracy, higher specificity).

Further analysis can be carried out to improve the predictions. For example, a Random Forest provides a lot of options for parameter tuning. Since training of a Random Forest is time consuming on a normal laptop, parameter tuning wasn't done in this study. Also, one could try to build other/more complex models. Finally, one could try to create new variables that might be useful in this classification problem.