CYO Adult Income

Felix Kleinpeter

8 2 2022

Contents

1 Executive Summary

This report, which is created as the “choose your own” project part of the HarvardX Data Science Capstone course, explores models that attempt to predict whether respondents from the 1994 US Census reported an income of more or less than US$50K. The final model, a classification tree based on ten of the predictors given in the data, achieves an overall accuracy of 85.29% with a sensitivity of 93.57% and a specificity of 61.44% when testing it against the validation set that is put aside before the model creation. This model fails to achieve the benchmark for success set at 85% overall accuracy with at least 75% sensitivity and specificity.

2 Introduction

Differences in income, and thus resulting wealth, have been subject of debates and investigations for a long time. Especially in countries with an inclination towards capitalism like the United States of America, the differences between the rich and the poor have also been growing over time. In the public discourse, this has often led to a discussion between different narratives: are the rich just harder-working, smarter or more fortunate than the poor? Or is there an ingrained bias towards people that exhibit certain characteristics that would make it easier for them to achieve a higher level of income?

In this report for the “Choose Your Own” project part of the HarvardX Data Science Capstone course, I examine data from the 1994 United States Census, made available on Kaggle. The data contains a sample of attributes as reported by respondents of the 1994 census, including information on whether the person has an income of more or less than US$50’000.

In my opinion at the outset of this project, as jobs (and the incomes that come with them) are not randomly assigned across the population, it should be possible to create a machine learning model that takes the given attributes of a person to predict their income with a reasonable accuracy. I set the overall level of accuracy to count as “reasonably accurate” at 85%. However, from prior knowledge about studying income and wealth distributions, I don’t think that the sample is going to be evenly split between people making more than $50K and those making less than $50k. Instead, the lower-income section is probably going to make up the majority of the population in the sample. In order to avoid being able to just “play the numbers” and guess a below-50K income disproportionately often, I include another stipulation in the definition of “reasonable accuracy”: Both sensitivity and specificity should exceed 70%. This provides a need to avoid wrongly attributing too many of the above-50K income class to the lower-income section.

3 Data Exploration

This chapter contains an overall first investigation of the given census data before diving deeper into looking at the individual predictors to evaluate their visually perceived helpfulness in building a decisioning model.

3.1 Data Setup and Investigation

Table 1: Raw data summary

age	workclass	fnlwgt	education	education.num
Min. :17.00	Length:32561	Min. : 12285	Length:32561	Min. : 1.00
1st Qu.:28.00	Class :character	1st Qu.: 117827	Class :character	1st Qu.: 9.00
Median :37.00	Mode :character	Median : 178356	Mode :character	Median :10.00
Mean :38.58	NA	Mean : 189778	NA	Mean :10.08
3rd Qu.:48.00	NA	3rd Qu.: 237051	NA	3rd Qu.:12.00
Max. :90.00	NA	Max. :1484705	NA	Max. :16.00

marital.status	occupation	relationship	race	sex
Length:32561	Length:32561	Length:32561	Length:32561	Length:32561
Class :character	Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character	Mode :character
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA

capital.gain	capital.loss	hours.per.week	native.country	income
Min. : 0	Min. : 0.0	Min. : 1.00	Length:32561	Length:32561
1st Qu.: 0	1st Qu.: 0.0	1st Qu.:40.00	Class :character	Class :character
Median : 0	Median : 0.0	Median :40.00	Mode :character	Mode :character
Mean : 1078	Mean : 87.3	Mean :40.44	NA	NA
3rd Qu.: 0	3rd Qu.: 0.0	3rd Qu.:45.00	NA	NA
Max. :99999	Max. :4356.0	Max. :99.00	NA	NA

The original data set from Kaggle is comprised of 32561 rows (i.e. individuals), with the observations split into 15 columns. The following table gives an overview over the columns:

Table 4: Columns in the data set

Column	Description
age	Age of a respondent
workclass	Employment type (e.g. private sector, government employee, self-employed, . . .)
fnlwgt	'Final Weight' - a summary statistic for people with similar socio-economic status
education	Highest achieved education level
education.num	Numerical representation of the 'Education' variable
marital.status	Marital status
occupation	Description of the individual's occupation
relationship	Description of the individual's relationship

Column	Description
race	Race
sex	Sex
capital.gain	Individual's capital gains
capital.loss	Individual's capital losses
hours.per.week	Work hours per week
native.country	Country of origin
income	Description if income is above or below $50k

To get a better impression of the data, let's have a look at the first rows of the table:

Table 5: First rows of the data

age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship
90	?	77053	HS-grad	9	Widowed	?	Not-in-family
82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family
66	?	186061	Some-college	10	Widowed	?	Unmarried
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried
41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child
34	Private	216864	HS-grad	9	Divorced	Other-service	Unmarried

race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
White	Female	0	4356	40	United-States	<=50K
White	Female	0	4356	18	United-States	<=50K
Black	Female	0	4356	40	United-States	<=50K
White	Female	0	3900	40	United-States	<=50K
White	Female	0	3900	40	United-States	<=50K
White	Female	0	3770	45	United-States	<=50K

This overview shows that there are rows in the data with question marks for certain values. To proceed with a clean data set, I remove the rows that have a question mark in any of the columns. This leaves the data set with 30162 rows, 92.6% of the original data.

Finally, I take the cleaned data set and partition it into two new subsets: the first one, adult_inc, is the set that is going to be used for data exploration as well as initial setup and testing of the modeling approaches; the second one, validation, is going to be used for the final hold out test. This is the test in which I see if the models created with the adult_inc data predict the level of income accurately on the until-then unknown validation data set. I choose the validation data set to be 15% of the total cleaned data set as this is large enough to perform a meaningful hold out test on later (having to predict multiple thousands of data points) while still giving a larger initial test set that allows to use techniques such as k-fold cross-validation.

3.2 Data Visualization

Having cleaned the data, I now investigate each parameter visually to find the predictors I want to include into my models later. Creating models based on a large number of predictors vastly increases the run time each model requires to make the prediction, so striking a balance between the number of predictors used and the marginal improvement in prediction is crucial to not optimize the model for theoretical, but practical use.

3.2.1 Age

As salaries usually are higher in positions that require more relevant work experience than in positions that can be picked up by people that are new to the workforce, I would expect there to be a distinct difference between the age distributions for respondents making more than and less than $50K. The majority of young people should probably be in the lower-income tier whereas the age groups between 40 and 60 years should be more evenly split between the income tiers. Finally, I would expect the balance to swing back in favor of the lower-income tier for senior citizens who have left the workforce and are relying on pensions for their income now.

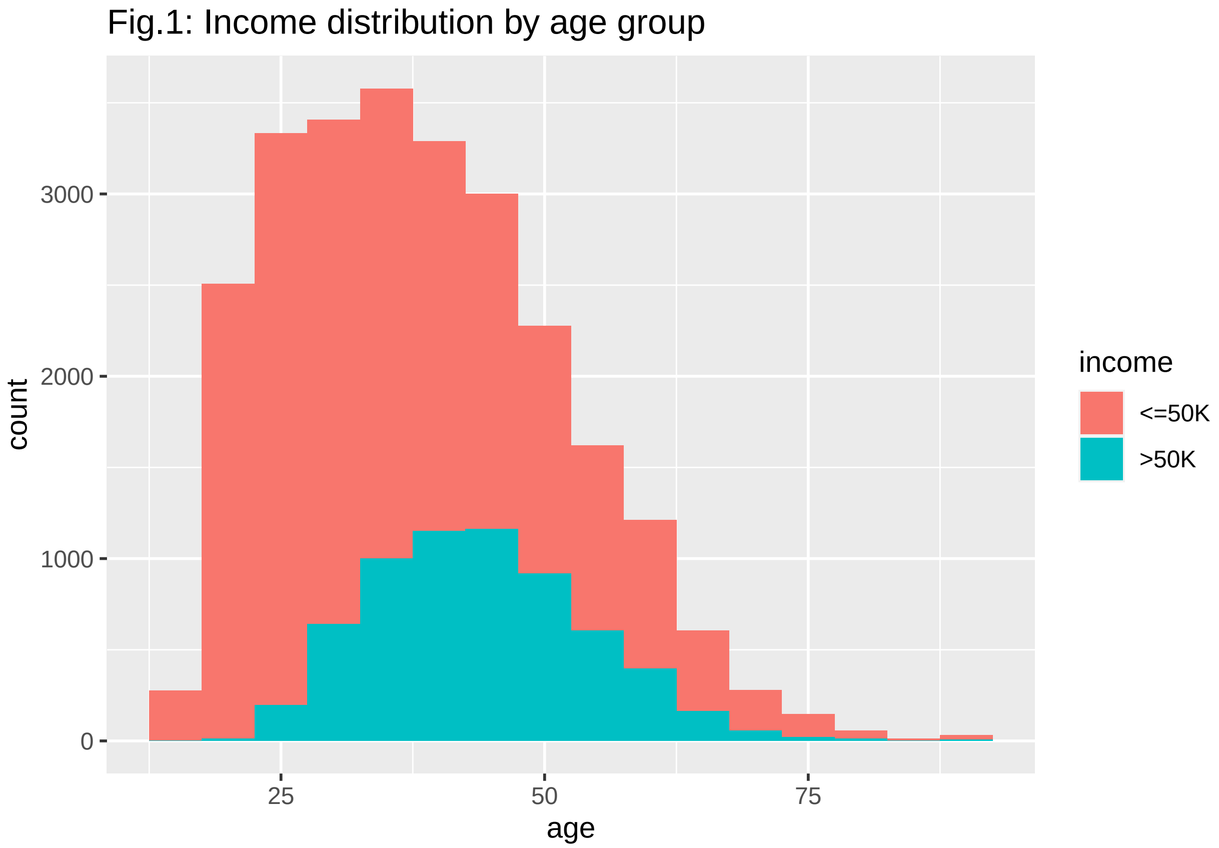
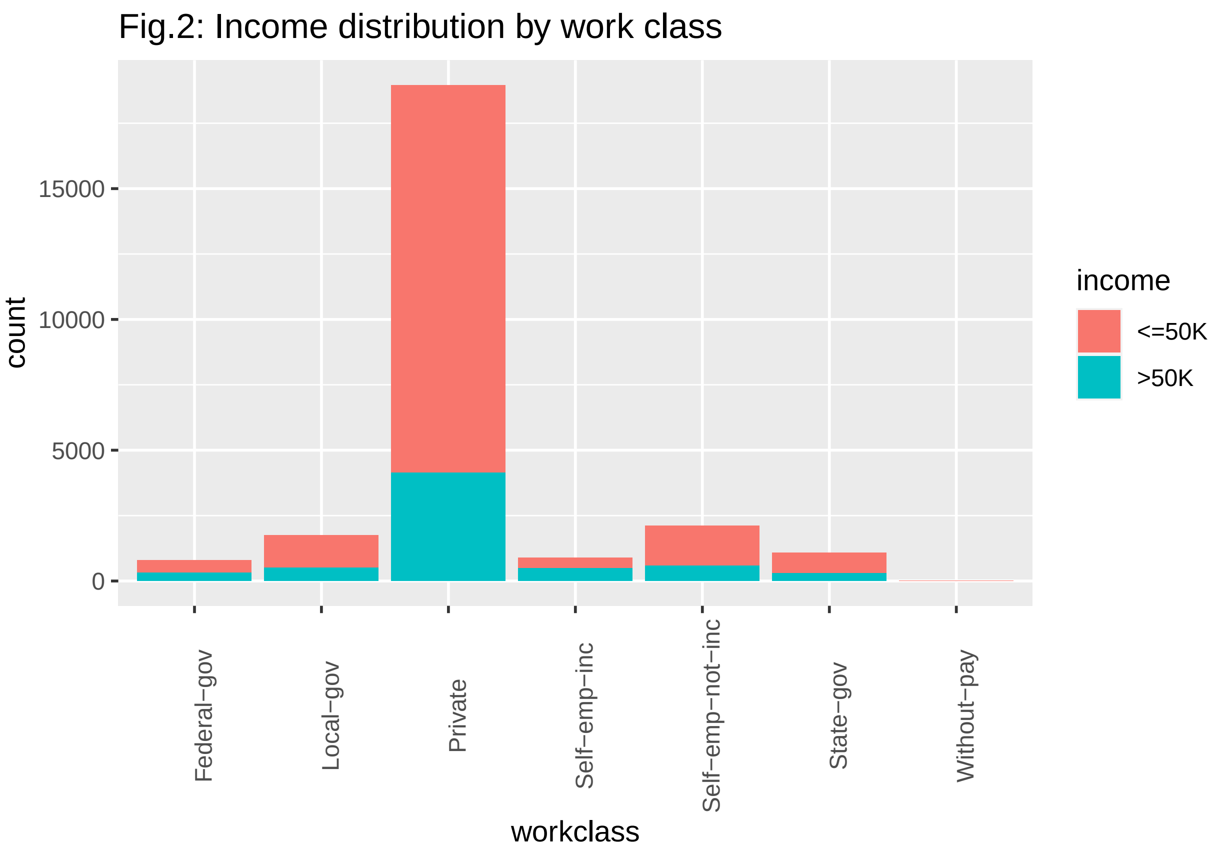


Fig. 1 indicates that the assumptions generally hold true. For middle-aged respondents, the relative share of high-income earners grows strongly compared to those who are either very young or very true. There seems to be a lot of sense in using age as one of the predictors.

3.2.2 Workclass

Often times, different employment sectors are seen as focusing on different priorities: jobs with in a government administration for example might pay a bit less, but provide higher job security than jobs in the private sector where the additional risk of losing the job is rewarded through a higher salary. Thus, there could again be a clear split between the ratios of high-income and low-income earners in the different work classes. Figure 2 investigates this.



While Fig. 2 shows that the majority of high-income earners are indeed employed in the private sector, this is mostly down to the overwhelming prevalence of private sector workers in the sample. Looking at the ratios between high- and low-income per sector, I find the differences between the types of government-based employment (Federal-gov, Local-gov and State-gov) as well as the difference between the different classes of self-employment (Self-emp-inc and Self-emp-not-inc) most interesting.

3.2.3 Fnlwgt

As the Fnlwgt column is a summary statistic already based on some of the predictors and the publishers of the data on Kaggle even state that "People with similar demographic characteristics should have similar weights. There is one important caveat to remember about this statement. That is that since the CPS sample is actually a collection of 51 state samples, each with its own probability of selection, the statement only applies within state.", I am not going to use Fnlwgt as a predictor in any of my models since the data does not include any information about the state the respondent lives in.

3.2.4 Education and Education.num

The data set contains two education-related columns: education, which describes the highest achieved education level in words, and education.num which seems to do the same, but assigns a number to the different education levels (the higher the number, the higher the level of education). To see if they actually match up, I will remove duplicate combinations of education and education.num to see if a number can be distinctly matched to an education level. The following table shows the results:

Table 7: Education levels in the data set

education	education.num
Preschool	1
1st-4th	2
5th-6th	3
7th-8th	4
9th	5
10th	6
11th	7
12th	8
HS-grad	9
Some-college	10
Assoc-voc	11
Assoc-acdm	12
Bachelors	13
Masters	14
Prof-school	15
Doctorate	16

As can be seen in the table, the education levels actually match the values of education.num with an increasing number in education.num indicating a higher degree of formal education. Given that higher education is generally sought after, but harder to obtain than just a high school degree, I assume that the ratio of the high earners should be strongly in favor for the higher levels of education while those with a lower level of income should mostly also fall into the low-income group. I investigate this in the following graph:

Fig. 3: Income levels by education

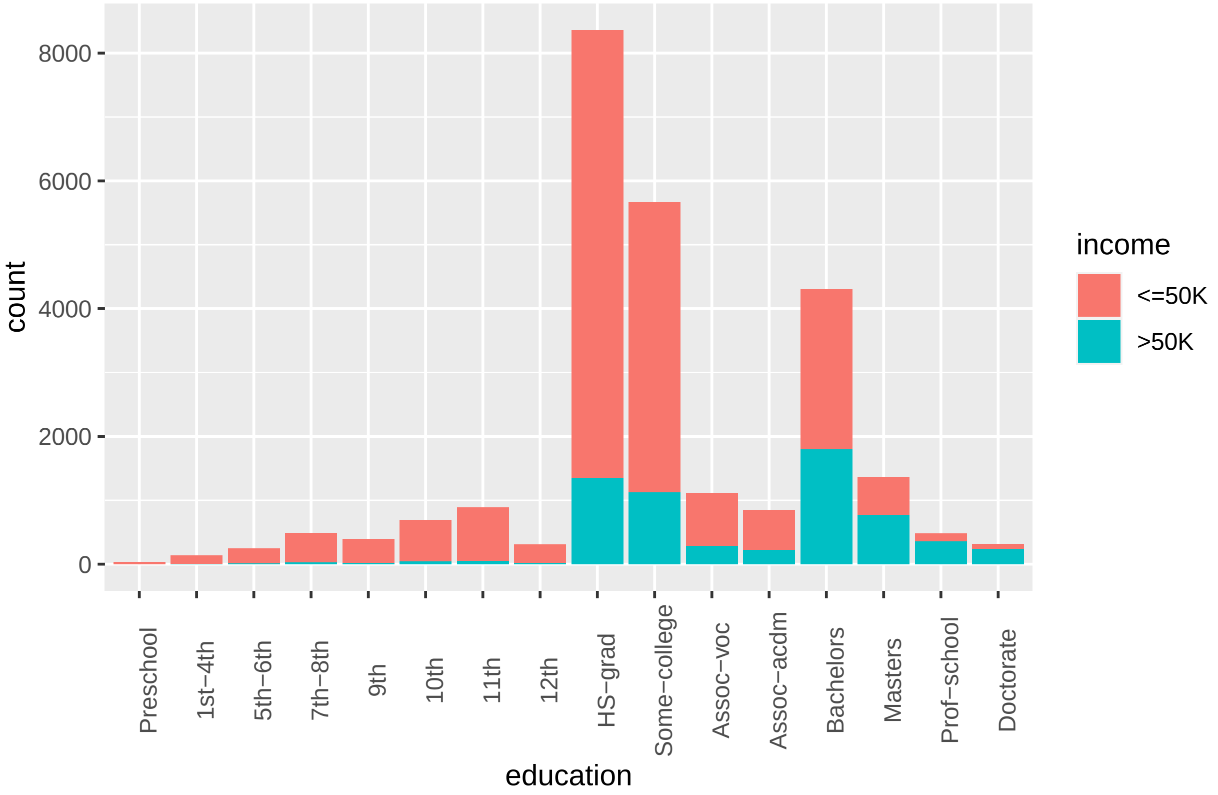


Fig. 3 confirms the suspicions that higher education levels correlate with higher incomes. Any degree below a Bachelors shows only a small fraction of respondents with an income above $50K. However, this ratio approaches almost half for workers with a Bachelors and a majority of workers with a Masters, Prof-school or Doctorate fall into the high-income class. However, as both predictors education and education.num express the same information, I will only use one of them in any model.

3.2.5 Marital.status and Relationship

The next logical pair of predictors are marital.status and relationship. To understand better what the differences between these predictors are and what they express, I plot the different characteristics that these predictors take against each other:

Fig. 4: Relationship and marital status types

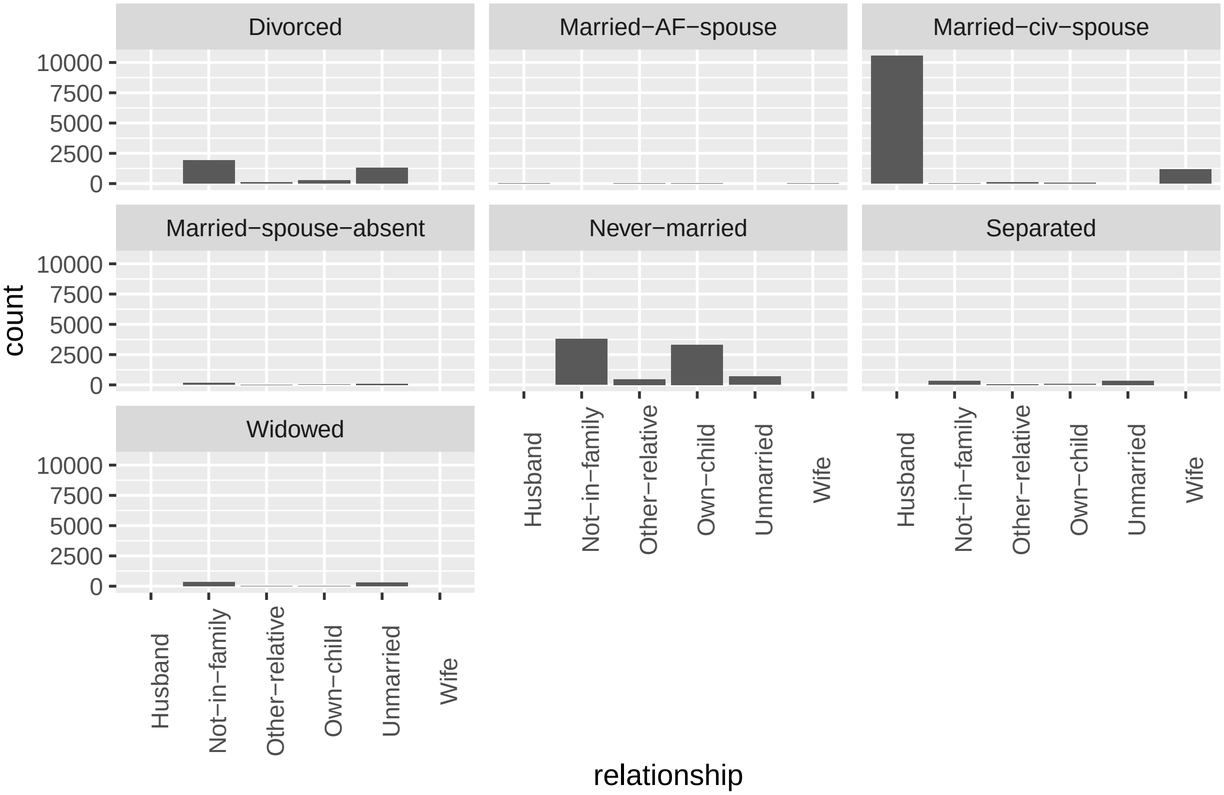
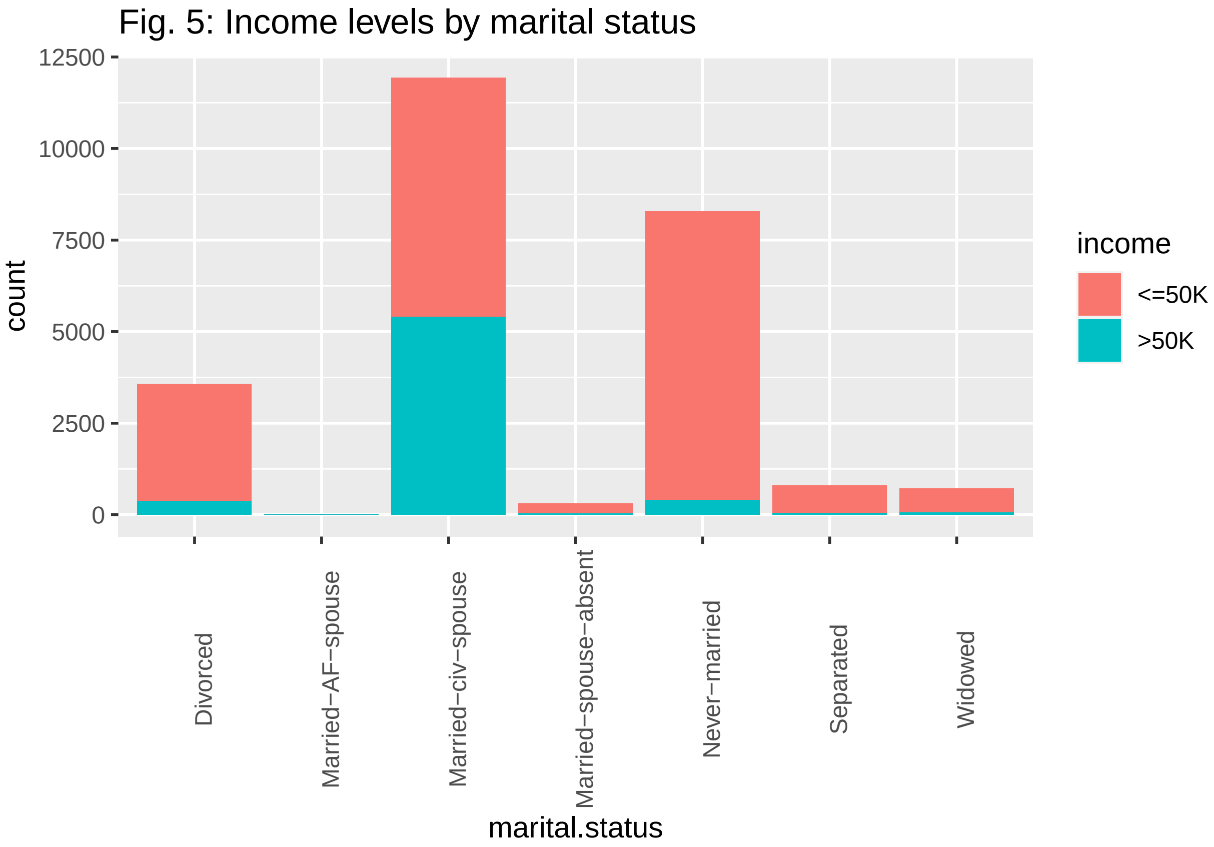


Fig. 4 shows that most of the respondents fall into the marital status types “Married-civ-spouse”, “Never-married” and “Divorced”. However, one thing stands out: the relationship option “unmarried” is available in noticeable proportions among “Divorced”, “Married-spouse-absent”, “Never-married”, “Separated” and “Widowed”. Looking at the various combinations of the marital.status and relationship predictors though, I find it very hard to then combine for example all “unmarried” respondents into one category, as the reasons for why they are unmarried can vary so broadly. Additionally, the “Husband” and “Wife” split contains information that could also be obtained via the Sex predictor. I therefore prefer to drop the relationship predictor from my models and focus on the marital.status instead. Fig. 5 will investigate if there might be a relationship between marital status and the income level.

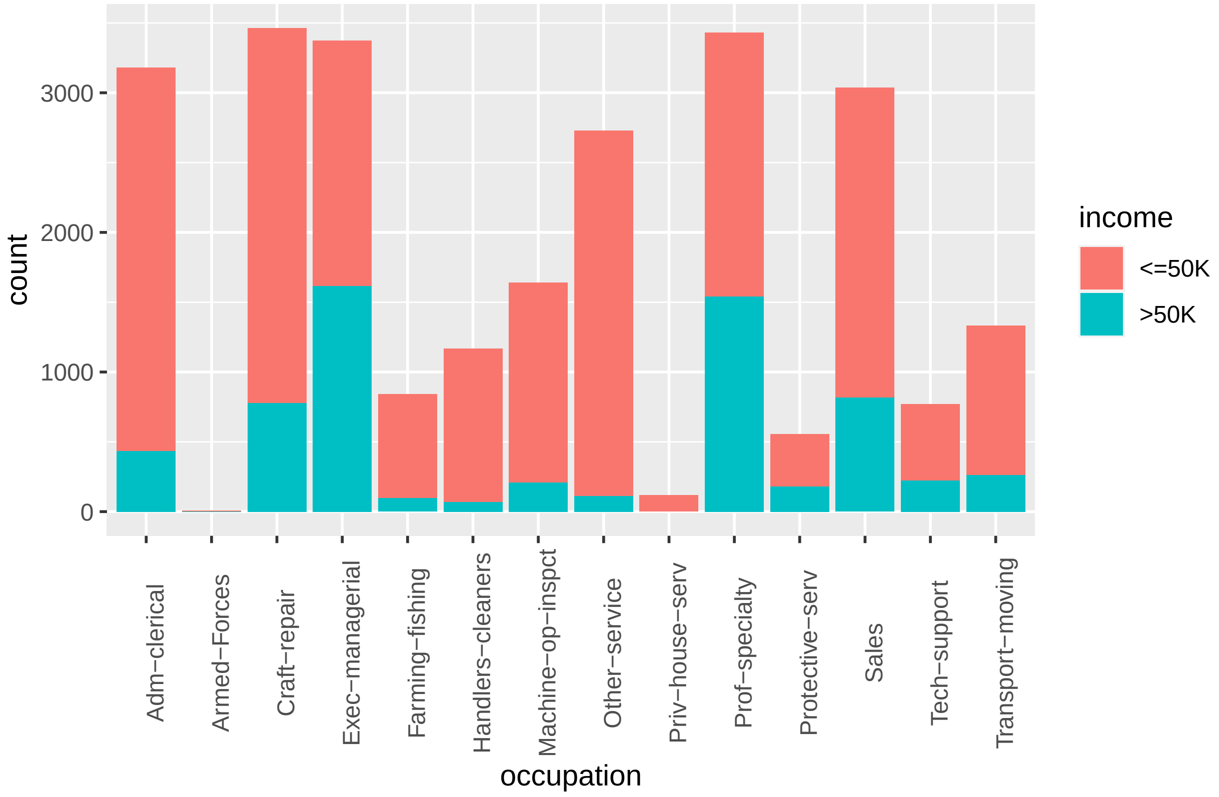


The plot shows very clearly that respondents that are marries have a much higher likelihood of being in the high-income group than those of any other marital status.

3.2.6 Occupation

Similar to the work class and age predictors earlier in the report, the occupation predictor is one that should show significant differences between the given characteristics. Some jobs (or job types) just pay significantly better than others and I expect the data to show this.

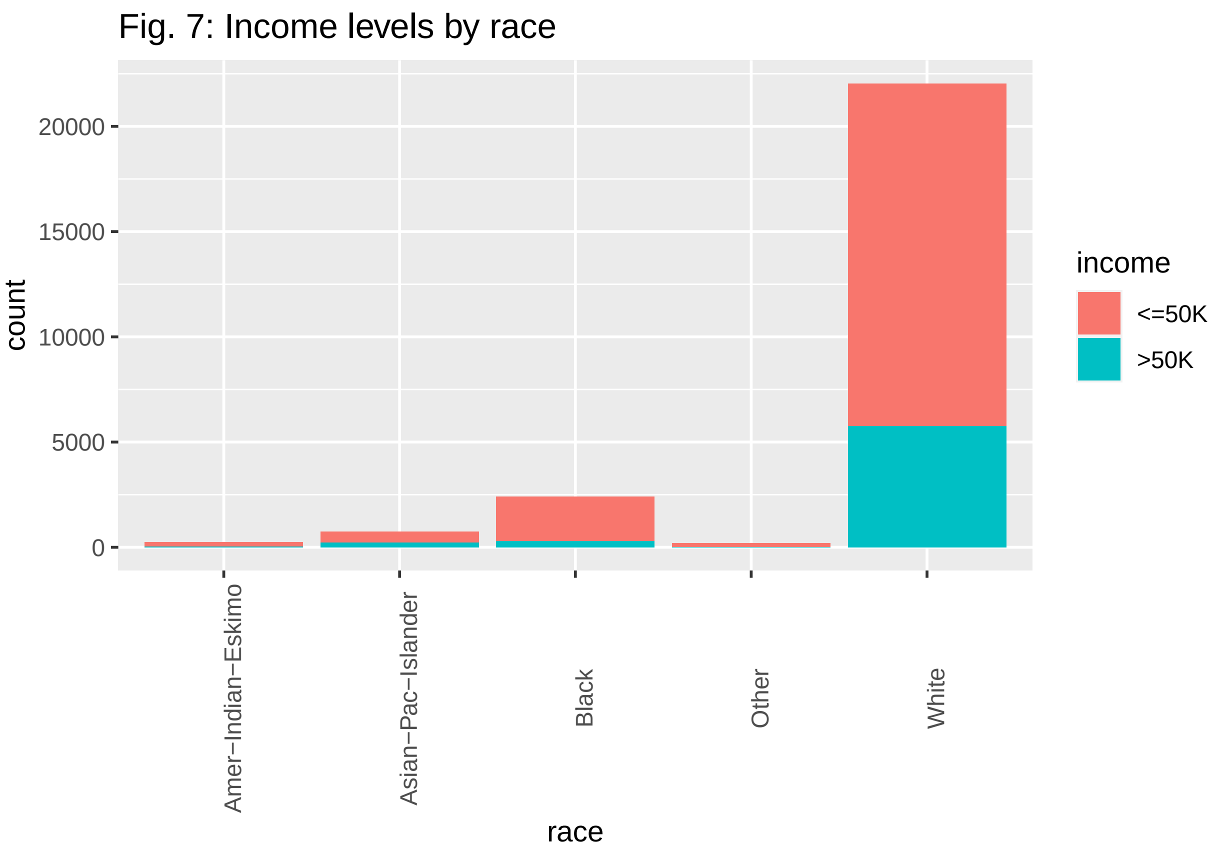
Fig. 6: Income levels by occupation



As expected, Fig. 6 confirms the assumptions. Respondents with occupations in the Exec-managerial or Prof-specialty fields have a much higher proportion of high-income respondents than those in the Handlers-cleaners or Other-service categories for example.

3.2.7 Race

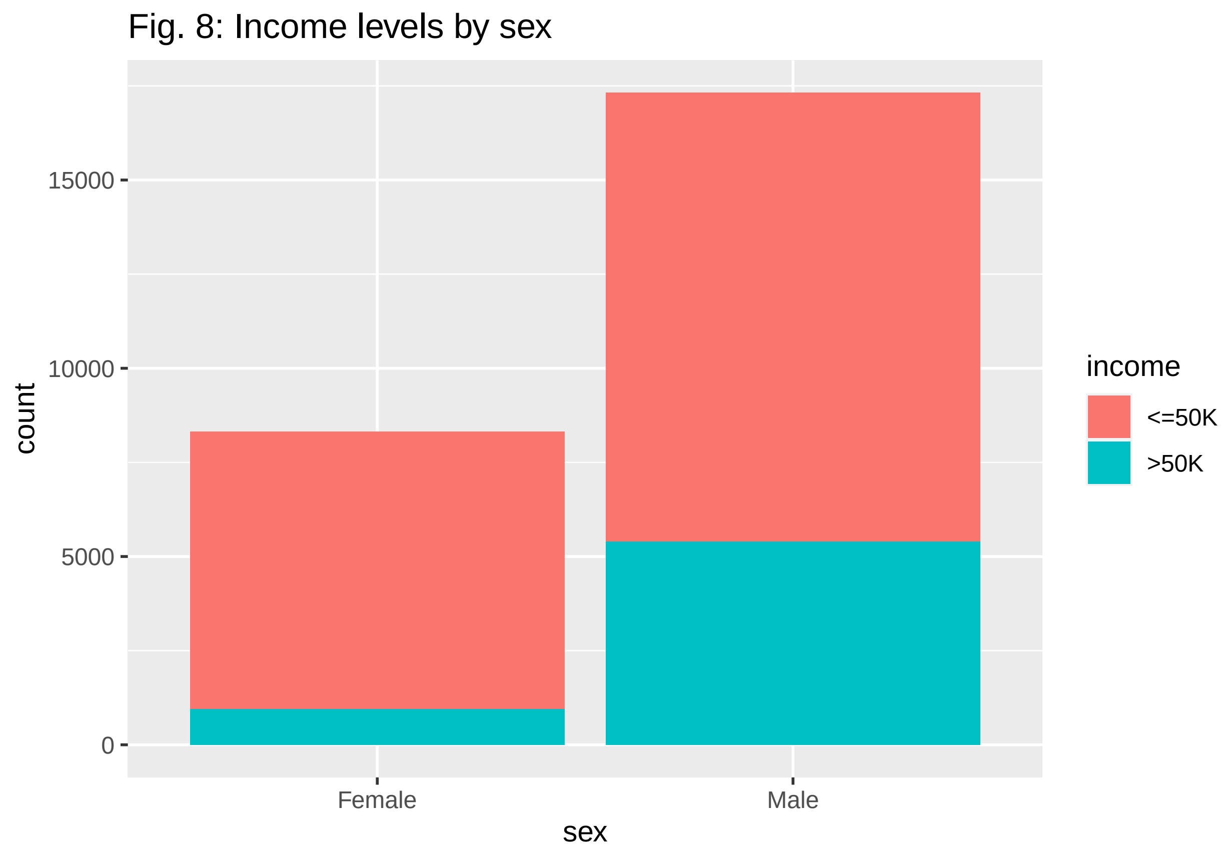
As the USA are an immigrant country (often times referred to as a melting pot of cultures and nations) and have experienced a lot of racial inequality along their history, it is interesting to see if race is still a defining factor for income levels in the given data from the 1994 census.



The visual inspection of Fig. 7 shows that the proportion of high-income seems to be highest among Whites and Asian-Pacific-Islanders, albeit the share of Whites in the sample being a lot larger at 22040 compared to the 758 Asian-Pacific-Islanders. The proportion of high-income earners drops off significantly for Blacks and Native Americans.

3.2.8 Sex

Similar to the impact of racial tensions in the development of somewhat equal incomes, sex is likely playing a role in the income distribution as well. I assume that given the "classical" roles in a family of a male breadwinner and female stay-at-home mother, the ratio of high-income earners should be clearly skewed towards males in comparison. Even to this day (as of 2022), there are differences in pay between males and females, so this effect is likely much stronger in data from 1994 that I am looking at here.

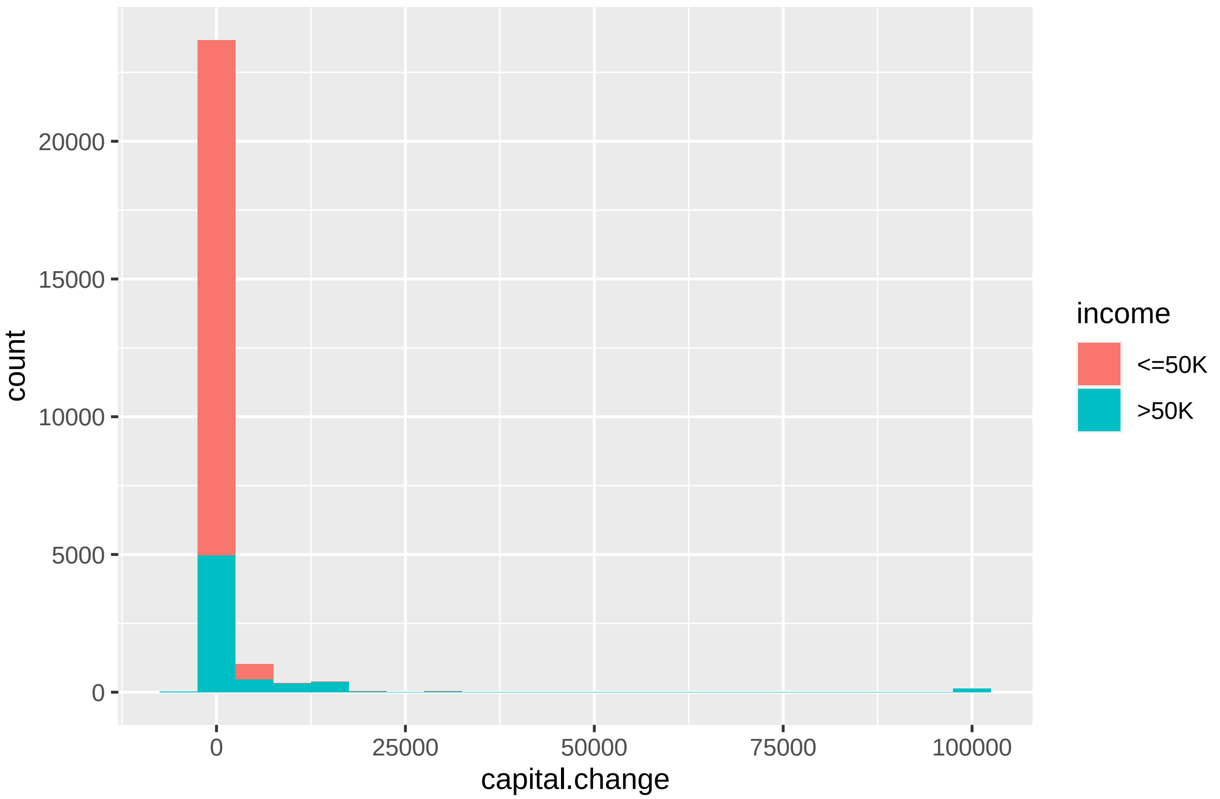


As expected, Fig. 8 shows a higher proportion of high-earners among the males than among females.

3.2.9 Capital.gain and Capital.loss

The data set gives information about capital gains and capital losses in the past year. To reduce the number of predictors for the final model, I combine the capital gains and losses into a new predictor "capital.change" which just subtracts the losses from any gains.

Fig. 9: Income levels by capital change



The graph shows that practically all of the respondents in the low-income category reported capital change that was close to zero whereas some of he higher-income respondents see capital changes in the multiple tens of thousands of dollars. Interestingly, the largest reported aggregated losses also come from respondents from the higher income group. This might be explained by the idea that lower-income respondents need all their income to pay for day-to-day expenses and don't have the ability to accrue larger losses as they don't have the resources to even invest an amount that would deliver such large losses.

3.2.10 Hours.per.week

A very straightforward argument for higher income, especially in today's "hustle culture", is the weekly amount of work hours. The argument goes that those who put in the work will also be rewarded by a higher income, thus there should be a clear split of those with higher and lower income across the work hour spectrum. Intuitively the split should sit somewhere around the typical 40-hour-week of a 9-to-5 job.

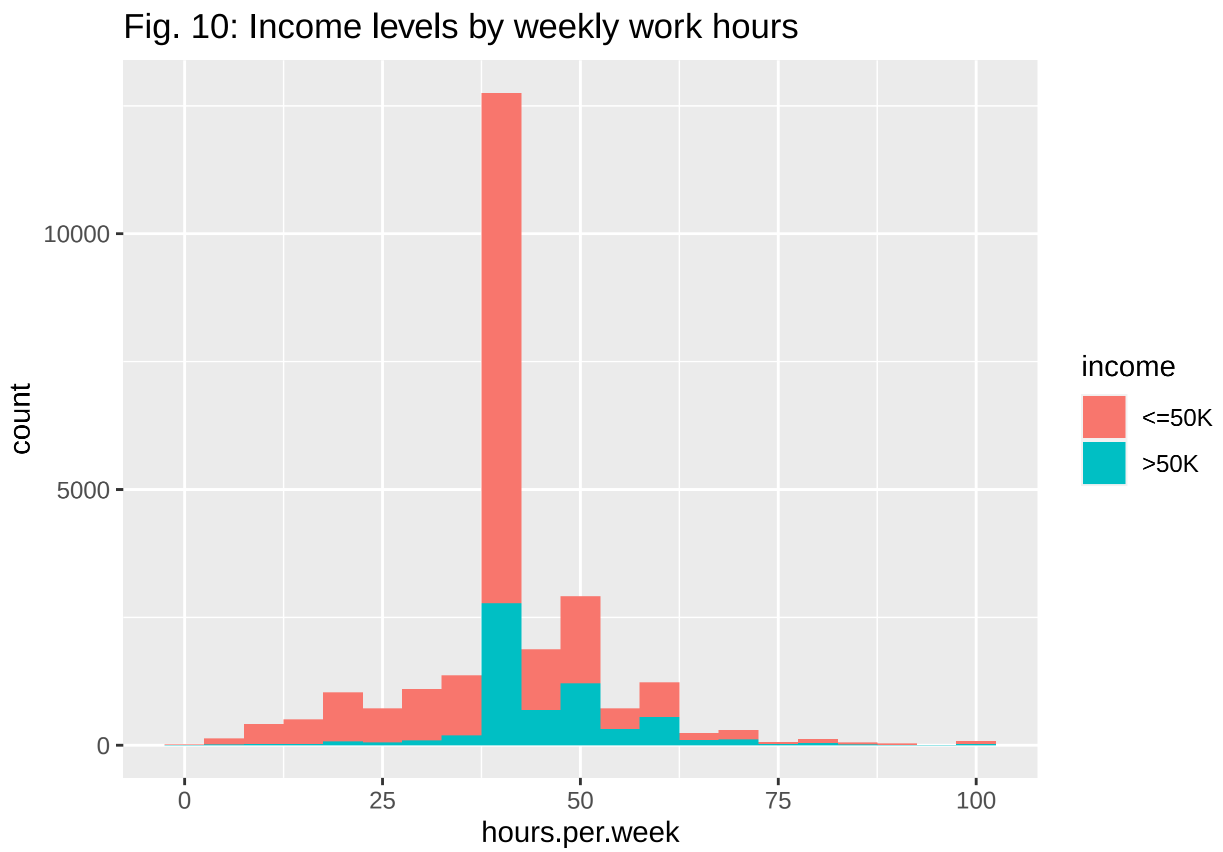


Fig. 10 shows that my assumption somewhat holds: the share of high-income earners drastically increases for those working 40+ hours per week and there are virtually no high-income earners who work less than this. However, for the people working more than 40 hours per week, there is still a significant amount of lower-income earners. They are still in a (albeit slimmer) majority for 50, 60 and 70 hours per week. This leads to the conclusion that while most of the higher income earners need to work at least 40 hours per week to make this salary, a high workload is certainly no guarantee for a high salary.

3.2.11 Native.country

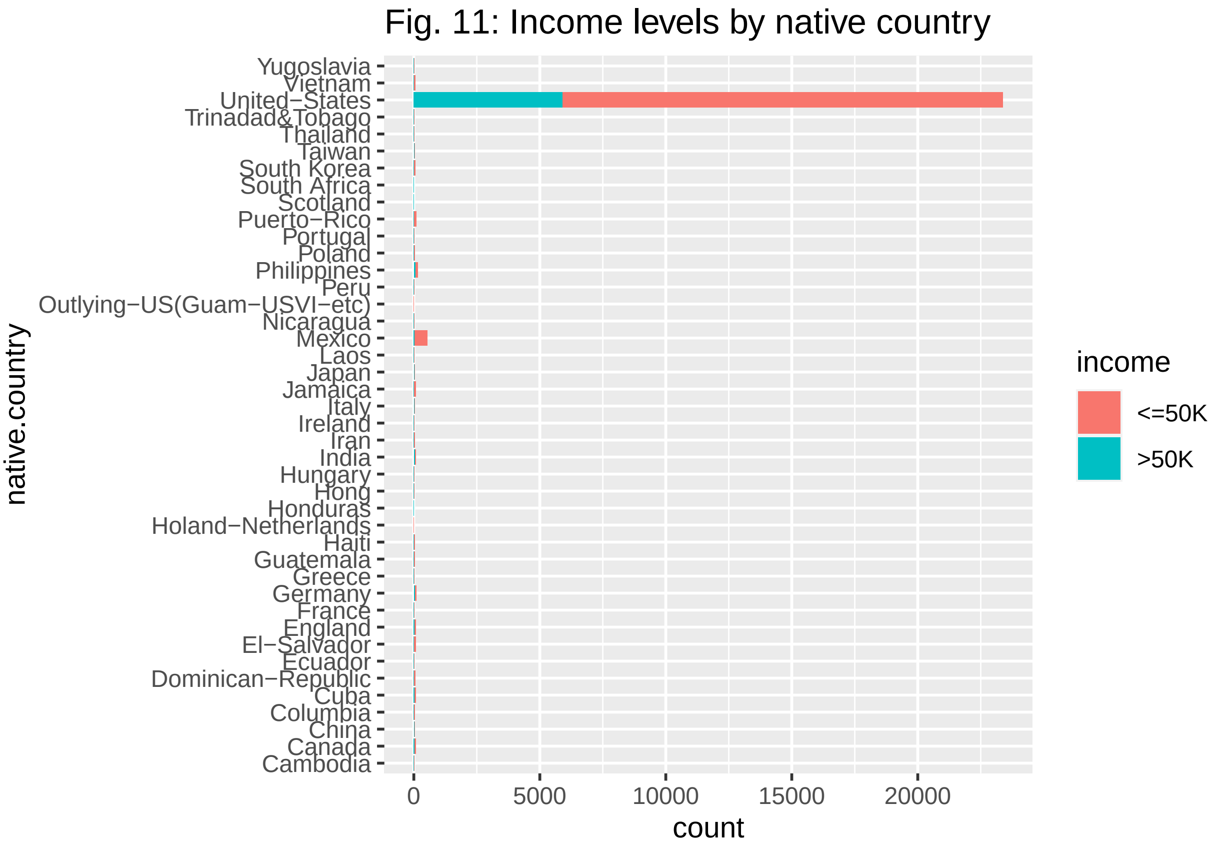
The last predictor in the data set is the native country. For this, let's first have a look at all the native countries in the sample:

Table 8: List of native countries in the sample

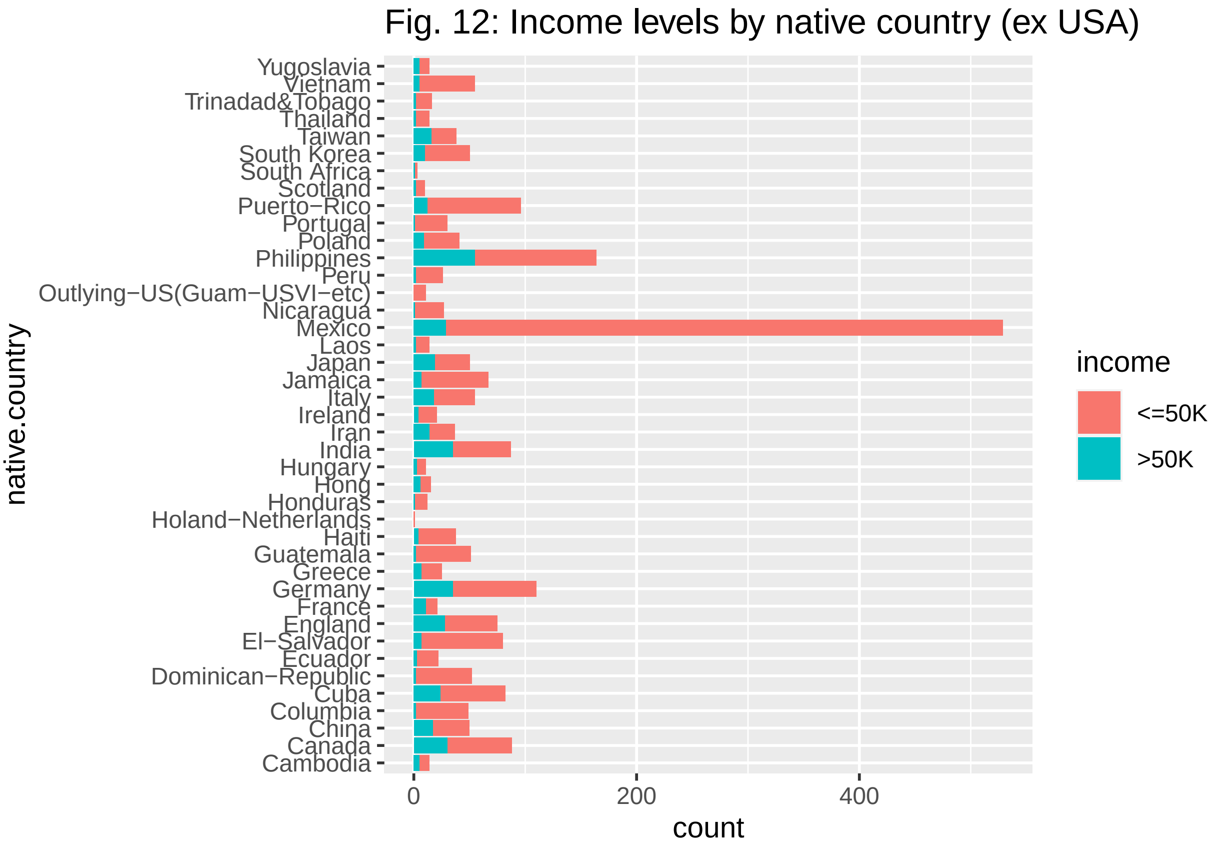
native.country	count
United-States	23385
Mexico	529
Philippines	164
Germany	110
Puerto-Rico	96
Canada	88
India	87
Cuba	82
El-Salvador	80
England	75

native.country	count
Jamaica	67
Italy	55
Vietnam	55
South	53
Dominican-Republic	52
Guatemala	51
China	50
Japan	50
Columbia	49
Poland	41
Haiti	38
Taiwan	38
Iran	37
Portugal	30
Nicaragua	27
Peru	26
Greece	25
Ecuador	22
France	21
Ireland	21
Trinadad&Tobago	16
Hong	15
Cambodia	14
Laos	14
Thailand	14
Yugoslavia	14
Honduras	12
Hungary	11
Outlying-US(Guam-USVI-etc)	11
Scotland	10
Holand-Netherlands	1

Immediately, it becomes clear that the vast majority of the respondents is from the United States, with Mexico following as the significantly largest country of origin for immigrants. There are two anomalies in the data, where the value for native.country isn't actually a full country name: "Hong", which can be easily resolved to be Hong Kong, and "South", which is more complicated as it could be either South Korea or South Africa. Given the demographics of these two countries, I assume that people from "South" with the race characteristic of "Asian-Pac-Islander" are from South Korea, while all other racial characteristics are from South Africa. With this settled, I can then plot the distribution per native country.



As shown in Fig. 11, the graph is dominated by the large number of respondents from the United States. To make the graph helpful at all, there are two possible ways: change the scale to a logarithmic rather than linear one or remove the United States natives from the graph to be able to compare the immigrants among themselves. I opt for the latter as there is one country (Holand-Netherlands) with just one respondent, which would not show up on a logarithmic scale.



On the scale of Fig. 12 it now becomes evident that there are clear differences between the native countries and the ratio of high-income respondents. Generally high income nations like France, England or Canada show much higher ratios of high-income respondents than countries such as Mexico, Guatemala, Haiti or the Dominican Republic. This might also be related to a geographical issue: while people from Central America and the Caribbean are already situated closer to the USA, they might be more tempted to move the relatively short way to get a better life for themselves, as even a low-paying job in the USA would be an improvement over the current situation. Compared to that, Europeans who move to the USA might go there to get an especially good education and would then have better chances to a high-paying job.

There are some noteworthy outliers though, such as a relatively high share of high-income earners that are originally from the Philippines or India.

4 Model Design

Having analyzed the data in the previous chapter, this part sets up multiple prediction models based on different modeling approaches. The adult_inc subset of the overall data will therefore be split into a test and a training set while the validation subset is still kept aside for the final hold out tests in the next chapter.

```r
# create test and train sets
set.seed(1910, sample.kind = "Rounding")
test_set_index <- createDataPartition(y = adult_inc$age, times = 1, p = 0.1, list = FALSE)
train_set <- adult_inc[-test_set_index,]
temp <- adult_inc[test_set_index,]

test_set <- temp %>%
  semi_join(adult_inc, by = "workclass") %>%
  semi_join(adult_inc, by = "education") %>%
  semi_join(adult_inc, by = "education.num") %>%
  semi_join(adult_inc, by = "marital.status") %>%
  semi_join(adult_inc, by = "occupation") %>%
  semi_join(adult_inc, by = "relationship") %>%
  semi_join(adult_inc, by = "race") %>%
  semi_join(adult_inc, by = "sex") %>%
  semi_join(adult_inc, by = "native.country")

removed <- anti_join(temp, test_set)

test_set <- rbind(test_set, removed)

# set up summary table
results_table <- tibble(Model = "Target",
                        Accuracy = 0.85,
                        Sensitivity = 0.75,
                        Specificity = 0.75)
```

4.1 Generalized Linear Model (GLM)

The first model I evaluate is a simple generalized linear model. Using the train()-function from R's caret package, I use age, workclass, education.num, marital.status, occupation, race, sex, capital.change, hours.per.week and native.country as predictors for this model.

```r
# train the model
train_glm <- train(income ~ age + workclass + education.num + marital.status +
                     occupation + race + sex + capital.change + hours.per.week +
                     native.country, method = "glm", data = train_set)
# create predictions for the test set
y_hat_glm <- predict(train_glm, test_set, type = "raw")
# create confusion matrix
cm_glm <- confusionMatrix(y_hat_glm, factor(test_set$income))
# add results to summary table
results_table <- bind_rows(results_table, tibble(Model = "GLM",
                                                 Accuracy = cm_glm$overall[[1]],
                                                 Sensitivity = cm_glm$byClass[[1]],
                                                 Specificity = cm_glm$byClass[[2]]))
```

As a first result, the relatively simple GLM model yields an accuracy of 84.0%, which is very close to the target of 85%. However, looking at the values for prevalence, sensitivity and specificity reveals how this accuracy was reached: Since the prevalence for the "<=50K" group is 75.6%, a high Sensitivity is going to have a strong impact on the overall accuracy. This is the case here, with a sensitivity of 91.9%, whereas the specificity is still relatively low at 59.5%.

4.2 Linear Discriminant Analysis

A model approach that builds upon the so-called naive Bayes model is Linear Discriminant Analysis. As the name suggests, this type of model splits the predicted outcomes into the different classes along a line. A downside of this is that due to the forced linearity of the model, it is unable to capture potential non-linearity in actual conditional probability functions. An advantage of this type of model however is that due to its relative simplicity, it allows usage of more parameters for the prediction. In this case, I'm estimating a model with the same predictors as in the Generalized Linear Model before.

```r
# train the model
train_lda <- train(income ~ age + workclass + education + marital.status +
                     occupation + race + sex + capital.change + hours.per.week +
                     native.country, method = "lda", data = train_set)
# create predictions for the test set
y_hat_lda <- predict(train_lda, test_set)
# create confusion matrix
cm_lda <- confusionMatrix(y_hat_lda, factor(test_set$income))
# add results to the summary table
results_table <- bind_rows(results_table, tibble(Model = "LDA",
                                                 Accuracy = cm_lda$overall[[1]],
                                                 Sensitivity = cm_lda$byClass[[1]],
                                                 Specificity = cm_lda$byClass[[2]]))
```

Again, the overall accuracy of 82.7% is very close to the target of 85%, but again this is carried mostly by the sensitivity value of 91.6%, rather than by the specificity which lags behind at only 55.2%.

4.3 Quadratic Discriminant Analysis

Another model approach that is in its origin very similar to the Linear Discriminant Analysis is the Quadratic Discriminant Analysis. The main difference here is that it can capture non-linearity, thus alleviating a main drawback of the Linear Discriminant Analysis. However, this comes at the cost of vastly increased complexity and computing time, making it less feasible to have as many parameters as in a Linear Discriminant Analysis. Additionally, too many parameters can quickly lead to overfitting for a QDA model. I thus reduce the number of predictors I use in my model. In this case, I choose age, education (represented by education.num), sex and marital status.

```r
# train the model
train_qda <- train(income ~ age + education.num + sex + marital.status,
                   method = "qda", data = train_set)
# create predictions for the test set
y_hat_qda <- predict(train_qda, test_set)
# create confusion matrix
cm_qda <- confusionMatrix(y_hat_qda, factor(test_set$income))
# add results to the summary table
results_table <- bind_rows(results_table, tibble(Model = "QDA",
                                                 Accuracy = cm_qda$overall[[1]],
                                                 Sensitivity = cm_qda$byClass[[1]],
                                                 Specificity = cm_qda$byClass[[2]]))
```

Interestingly, the overall accuracy of this model drops by a substantial amount compared to the others, showing a value of only 70.6%. This seems to be due to a much lower sensitivity than the other models produced, coming in at only 65.9%. Contrary to the previous models, the QDA model overachieves on the target specificity though, coming in at 85.0%. As the prevalence of the “positive class” of low income respondents is relatively high though, this increase in specificity can’t make up for the losses in the sensitivity measure compared to the other models.

4.4 Classification Tree

A more intuitively approachable model is a classification or decision tree. In this type of model, the predictors are partitioned to predict an outcome in essentially a yes-or-no question style. This makes it very easily understandable for anyone looking at the final model as it can generally be described as a flowchart of decision nodes. A change for this model compared to the previous ones is that it contains a so-called tuning parameter, the complexity parameter cp. By assigning multiple values to the tuning parameter, the model can be optimized for the best results. To use cross-validation and In our case, a classification tree model using the caret package’s rpart model looks like this:

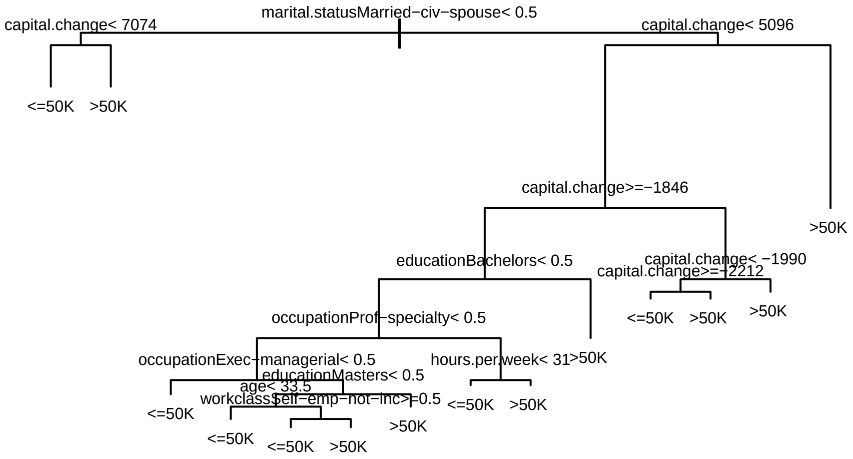
```
# train the model
train_rpart <- train(income ~ age + workclass + education + marital.status + occupation +
                       race + sex + capital.change + hours.per.week + native.country,
                     method = "rpart", tuneGrid = data.frame(cp = seq(0, 0.1, len = 25)),
                     data = train_set)
```

```
## Warning in (function (kind = NULL, normal.kind = NULL, sample.kind = NULL) :
## non-uniform 'Rounding' sampler used
```

```
# create predictions for the test set
y_hat_rpart <- predict(train_rpart, test_set)
# create confusion matrix
cm_rpart <- confusionMatrix(y_hat_rpart, factor(test_set$income))
# add results to results table
results_table <- bind_rows(results_table, tibble(Model = "Classification Tree (rpart)",
                                                 Accuracy = cm_rpart$overall[[1]],
                                                 Sensitivity = cm_rpart$byClass[[1]],
                                                 Specificity = cm_rpart$byClass[[2]]))
# best tune parameter
best_cp_rpart <- train_rpart$bestTune[[1]]
```

The decision tree achieves the highest accuracy of all models so far, coming in at 84.4%. This is again mostly due to the excellent Sensitivity of the model which is 92.9%. The specificity is again relatively low at 58.1%, just between the GLM and LDA models. The big advantage of the Decision Tree model is the mentioned easy interpretability. In this case, I can plot the decision rules at each node with the following code:

```
# plot the decision tree
plot(train_rpart$finalModel, margin = 0.1)
text(train_rpart$finalModel, cex = 0.5)
```



The chosen tuning parameter for this model was cp = 0.0041667.

4.5 Models summary

None of the models presented in this section reached the target of 85% accuracy in line with 75% sensitivity and specificity each. However, the exploration of the models has impressively shown the difference in strengths of some of the models. A particular outlier here was the QDA model which used a reduced number of predictors, but with those was able to achieve the by far highest specificity value - at the cost of a reduced sensitivity though, which due to the high prevalence also dragged down the overall accuracy.

Table 9: Accuracy, Sensitivity and Specificity of the explored models

Model	Accuracy	Sensitivity	Specificity
Target	0.8500000	0.7500000	0.7500000
GLM	0.8397661	0.9189886	0.5948963
LDA	0.8272904	0.9164087	0.5518341
QDA	0.7056530	0.6589267	0.8500797
Classification Tree (rpart)	0.8440546	0.9293086	0.5805423

5 Final Model Evaluation

Before the final hold out test can be run, I need to apply the same data transformations to the validation data set that I have applied to the adult_inc set during the data exploration.

```
# add capital.change
validation <- validation %>% mutate(capital.change = capital.gain - capital.loss)
# sort any native.country = "South" into South Korea and South Africa
validation <- validation %>%
  mutate(native.country = ifelse(native.country == "South" & race == "Asian-Pac-Islander",
                                "South Korea", native.country)) %>%
  mutate(native.country = ifelse(native.country == "South", "South Africa", native.country))
```

For the final model evaluation, I choose the Classification Tree (rpart) to see if this will perform well on the validation as well or if an amount of overfitting or overtraining has taken place in the model training.

```
# train model
final_train_rpart <- train(income ~ age + workclass + education + marital.status +
                             occupation + race + sex + capital.change + hours.per.week +
                             native.country, method = "rpart", cp = best_cp_rpart,
                           data = adult_inc)
# create predictions on validation set
final_y_hat_rpart <- predict(final_train_rpart, validation)
# create confusion matrix
cm_final_rpart <- confusionMatrix(final_y_hat_rpart, factor(validation$income))
# final results table
final_table <- tibble(Model = c("Target", "Final Classification Tree (rpart)"),
                      Accuracy = c(0.85, cm_final_rpart$overall[[1]]),
                      Sensitivity = c(0.75, cm_final_rpart$byClass[[1]]),
                      Specificity = c(0.75, cm_final_rpart$byClass[[2]]))
knitr::kable(final_table, caption = "Final Model Results")
```

Table 10: Final Model Results

Model	Accuracy	Sensitivity	Specificity
Target	0.8500000	0.750000	0.750000
Final Classification Tree (rpart)	0.8011489	0.994939	0.243359

As the above table shows, the Classification Tree (rpart) model with the optimized complexity parameter from the model setup phase is able to improve the overall accuracy as well as both sensitivity and specificity when compared to the results from the tests with only the adult_inc data. However, the improvements in specificity are still not sufficient to achieve the target of 75% specificity that was set out in the beginning. The high accuracy thus still relies at least in part on a relatively high prevalence (74.2%) of low-income respondents in the validation data set.

6 Conclusion

The project managed to deliver a prediction model that fulfilled the criteria of an overall prediction accuracy above 85% and a sensitivity above 75%. However, the model failed to achieve the targeted specificity of 75% by a rather significant margin. Test with different types of models found that a classification tree seems to be the best type of model for predictions in this data set.

Of course, this work can be seen as only preliminary and could be expanded upon. For example, a general limitation of the way the data was gathered is that respondents to the 1994 census might have not fully understood what the applicable answer for any question for their life situation is, or just purposefully answered wrongly in order to mislead the researchers. Some form of data validation could help to alleviate this situation. Additionally, the type of prediction models used could be expanded. Since the classification tree appears to be the dominant type of model here, exploring multiple of those via a so-called random forest could be at the heart of a follow up project. Alternatively, an ensemble model between the QDA (which produced a high specificity value) and the classification tree could help fix the low specificity achieved in the final model.