Income based on US Census data
Exploratory Data Analysis

The distribution of our target variable “income” is shown below on Fig 1. 76% are people with income less or equal than $50k, the remaining 24% have income greater than $50k.

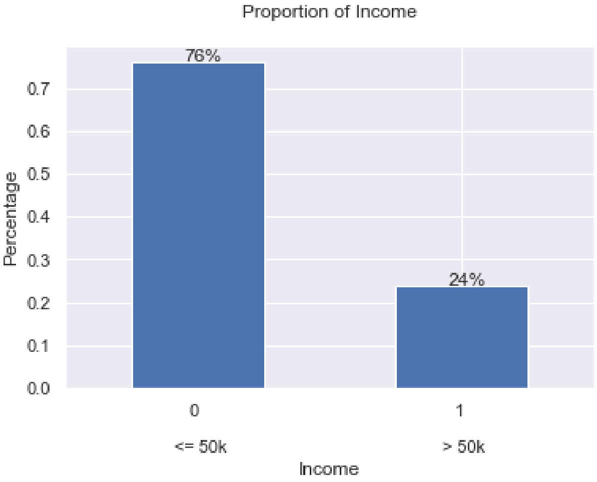
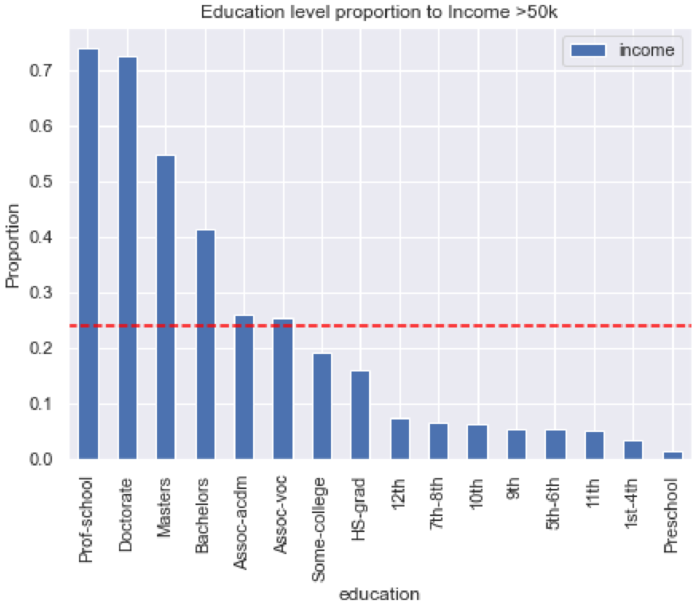
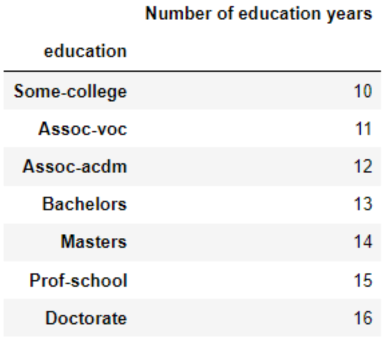


Fig. 1

Some of the main factors that could contribute to an individual’s income can be level of education, industry, location, work type, age and gender. We’ll explore the ones included in our dataset.

Usually a high level of education could result in a job that is well paid. Our data confirms our assumption, Fig. 2 shows Professional school, Bachelors, Masters and Doctorate as having the highest probability of income greater than $50k. It’s worth mentioning that Professional school ranks above the other 3 and in this context corresponds to what it seems an education level higher than Masters.

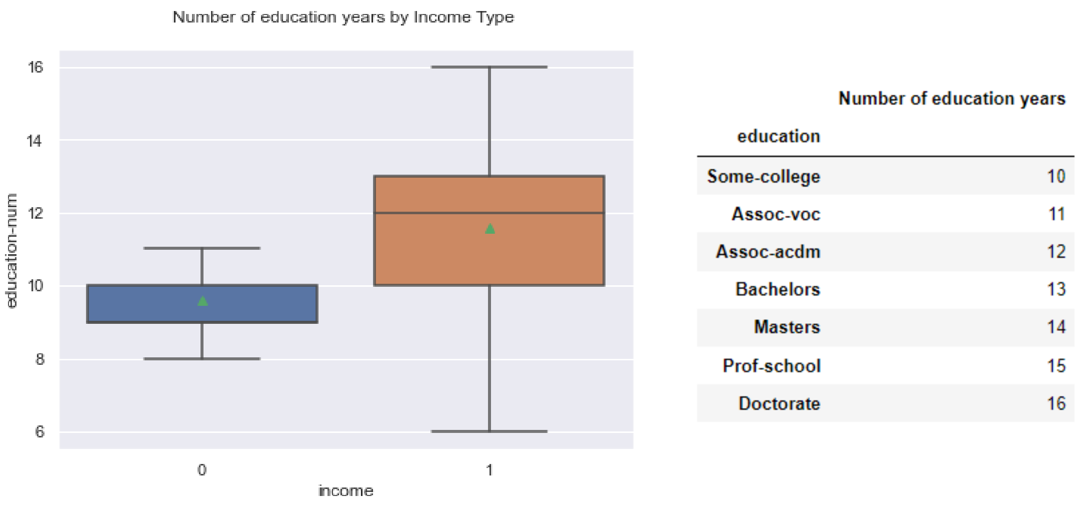




	Number of education years
education	
Some-college	10
Assoc-voc	11
Assoc-acdm	12
Bachelors	13
Masters	14
Prof-school	15
Doctorate	16

Fig. 2

We also can take a look at the number of years of education on Fig 3. About 75% of people with income greater than $50k have at least 10 years of education, which corresponds to at least some college.

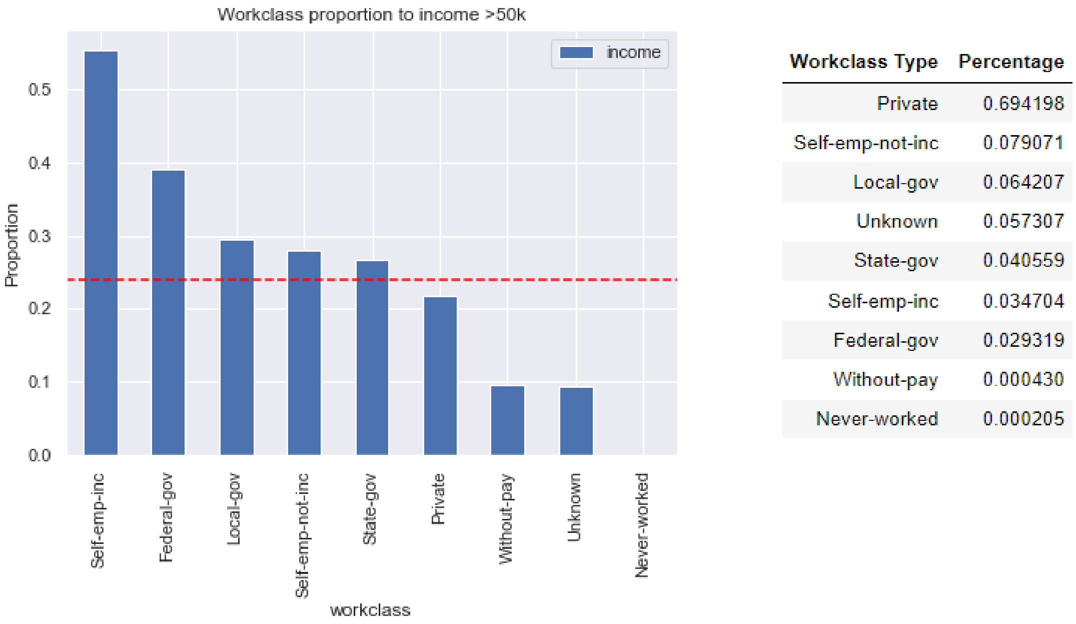


education	Number of education years
Some-college	10
Assoc-voc	11
Assoc-acdm	12
Bachelors	13
Masters	14
Prof-school	15
Doctorate	16

Fig. 3

Other factors we had mentioned at the beginning are industry, location and work type. Unfortunately we don't have location nor industry data. Regarding work type we can look at workclass and occupation.

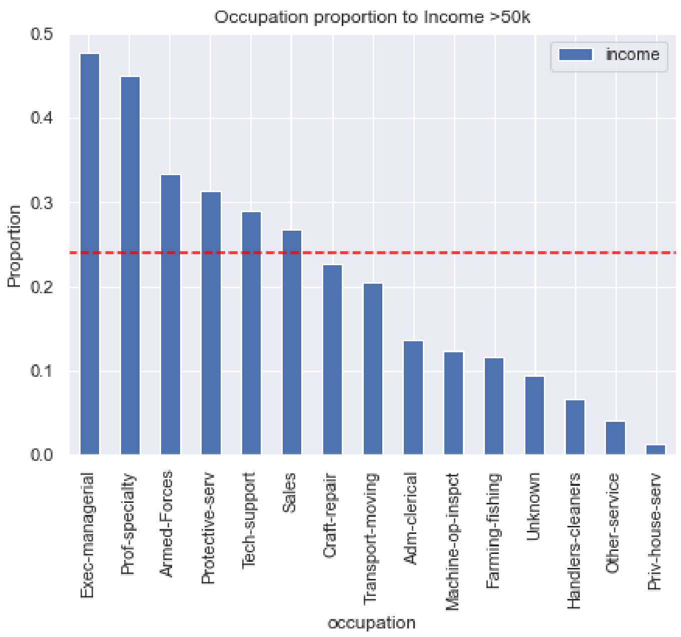
Fig. 4 shows us workclass distribution. Self employed (Inc) seems to have the highest probability of higher income, but this category only accounts for 3.5% of our data. The private sector accounts for almost 70% of our data and it appears to lean more towards the income being less than $50k.

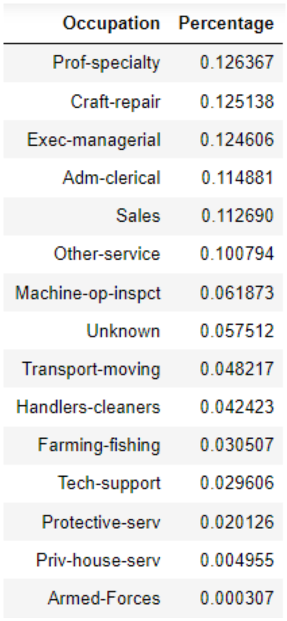


Workclass Type	Percentage
Private	0.694198
Self-emp-not-inc	0.079071
Local-gov	0.064207
Unknown	0.057307
State-gov	0.040559
Self-emp-inc	0.034704
Federal-gov	0.029319
Without-pay	0.000430
Never-worked	0.000205

Fig. 4

Following our exploration of work type, Fig 5 shows us the occupation breakdown. As expected, Executives, Management and Professionals are the occupations with higher probability of income greater than $50k and they account for 25% of our data. Armed Forces and Protective Services follow, but they only account for 2%, so not very significant.



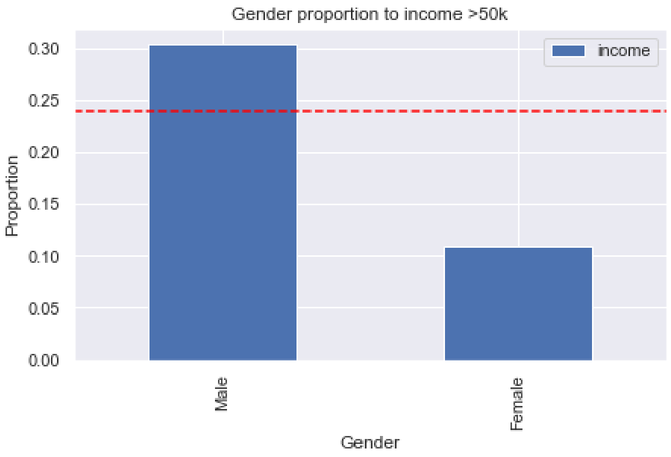


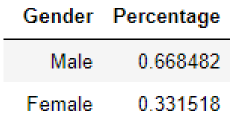
Occupation	Percentage
Prof-specialty	0.126367
Craft-repair	0.125138
Exec-managerial	0.124606
Adm-clerical	0.114881
Sales	0.112690
Other-service	0.100794
Machine-op-inspct	0.061873
Unknown	0.057512
Transport-moving	0.048217
Handlers-cleaners	0.042423
Farming-fishing	0.030507
Tech-support	0.029606
Protective-serv	0.020126
Priv-house-serv	0.004955
Armed-Forces	0.000307

Fig. 5

Let’s explore our remaining factors from the initial list: gender and age.

Fig 6 shows a strong differentiation in income due to gender. Being male almost triples the probability of higher income over being female. These data from 1994 still represent a more favorable income if the individual is a male. Our data also show 67% of the census respondents were males and 23% were females.

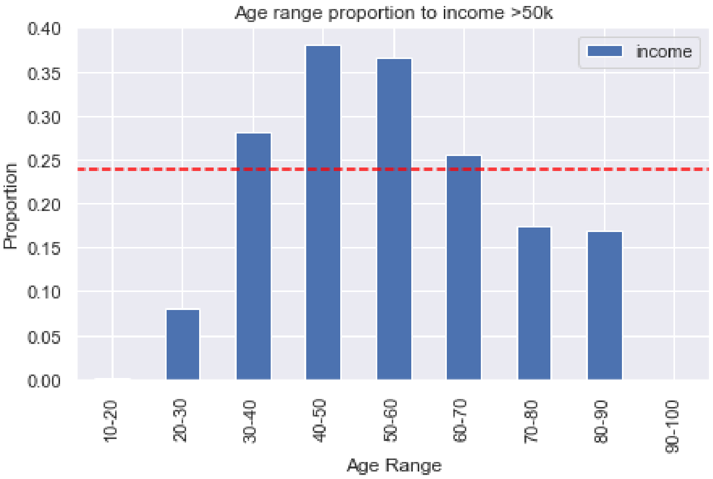


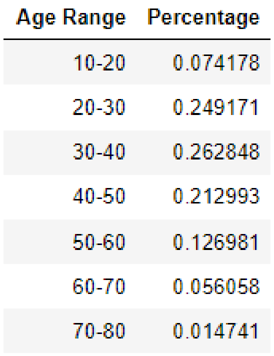


Gender	Percentage
Male	0.668482
Female	0.331518

Fig. 6

Finally let's explore if age is relevant for higher income. Fig. 7 shows higher probabilities between ages 40 to 60. We then see a decline after 60 years, which could be related to people's retirement.



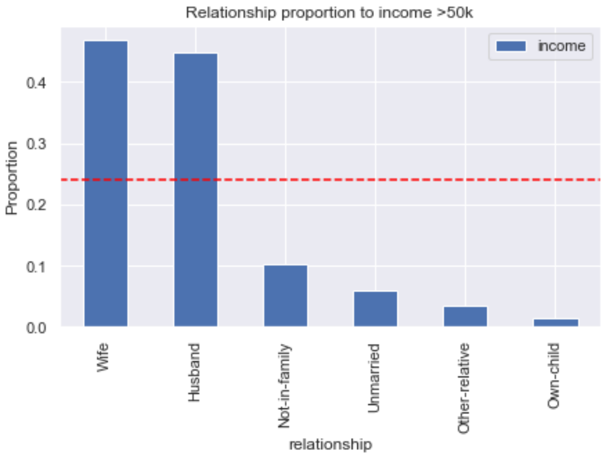


Age Range	Percentage
10-20	0.074178
20-30	0.249171
30-40	0.262848
40-50	0.212993
50-60	0.126981
60-70	0.056058
70-80	0.014741

Fig. 7

During our data analysis, we found other factors that could contribute to high income: marital status/relationship, the amount of hours worked per week and capital gains from investments.

Both Fig. 8 and Fig. 9 show that you have a better probability of high income if you're married.



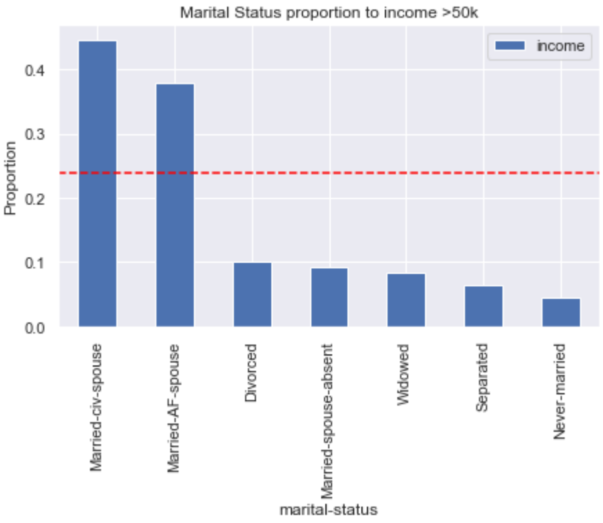


Fig. 8

Fig. 9

Hours worked per week also show a differentiation between high income or not (Fig. 10). People who work more than the standard 40 hrs per week have a better probability of high income. According to this article, earnings rise as a function of hours worked up to the 55 hours per week mark.

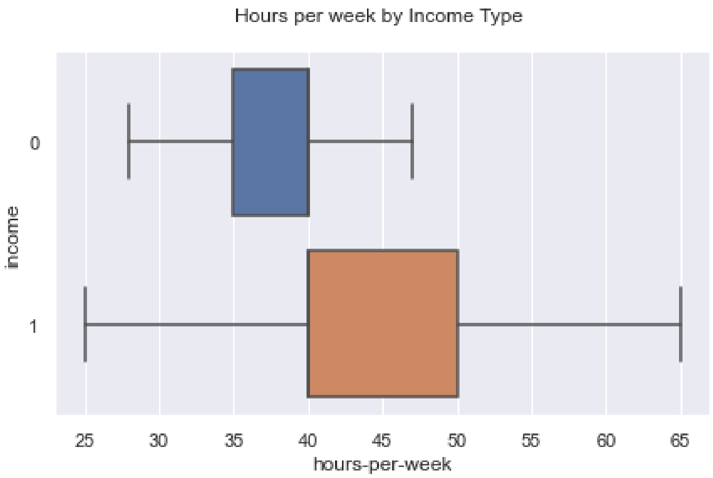


Fig. 10

Capital gains chart on Fig 11 shows that the majority of people with incomes greater than $50k have capital gains of $7,000 or more. Although this is a clear indicator, only 9% of people in our data have capital gains.

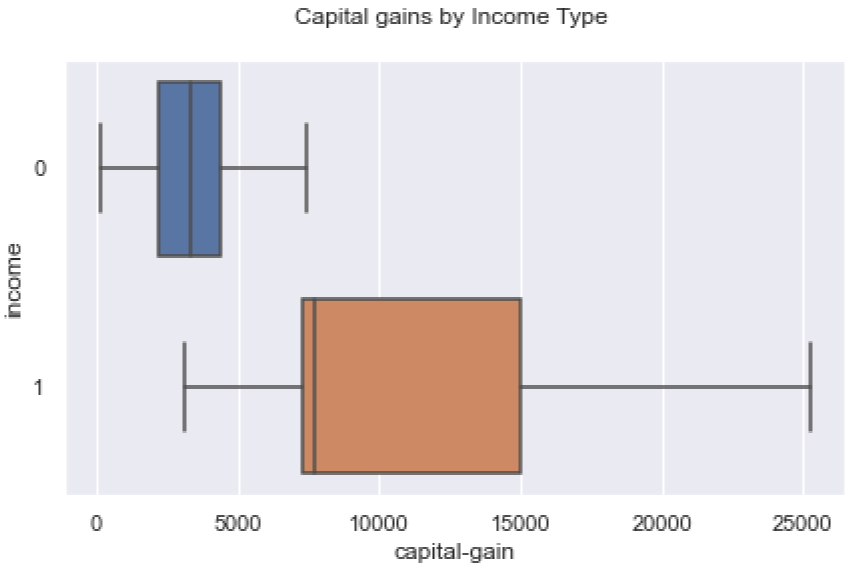


Fig. 11

The following correlation heatmaps in Fig 12 and 13 confirm some of the factors we found to influence higher income, such as level of education, age, occupation, marital status and gender.

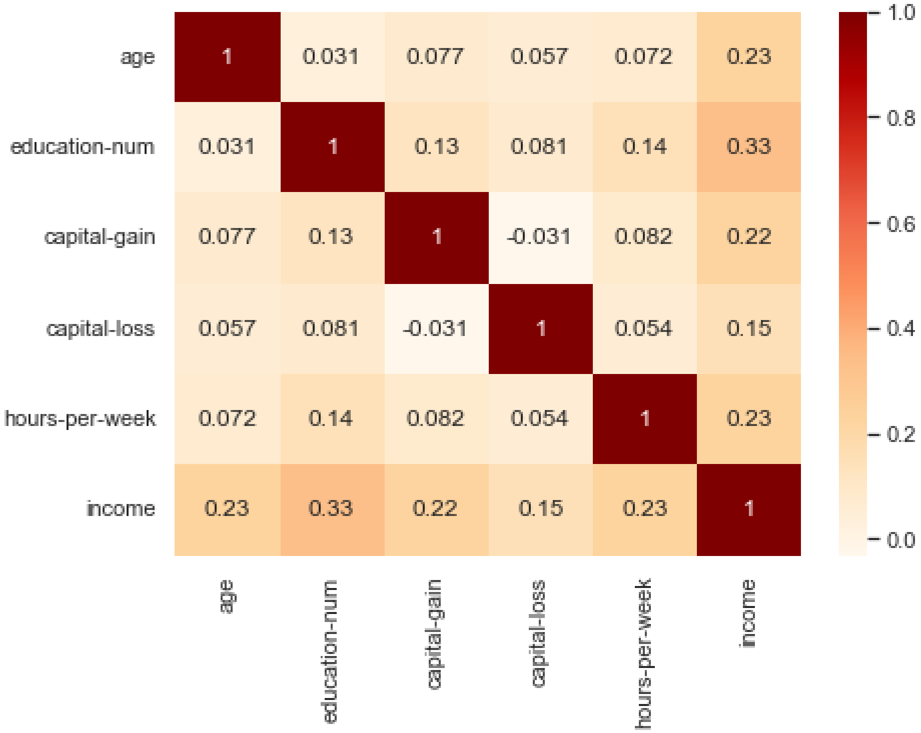


Fig. 12

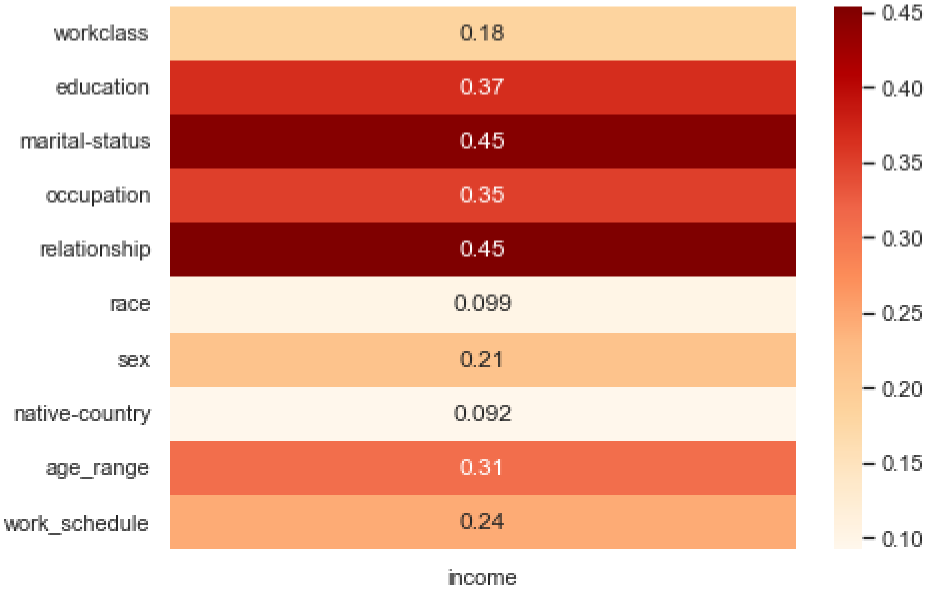


Fig. 13

Finally we built a basic Random Forest model to obtain the features importances, with the following results:

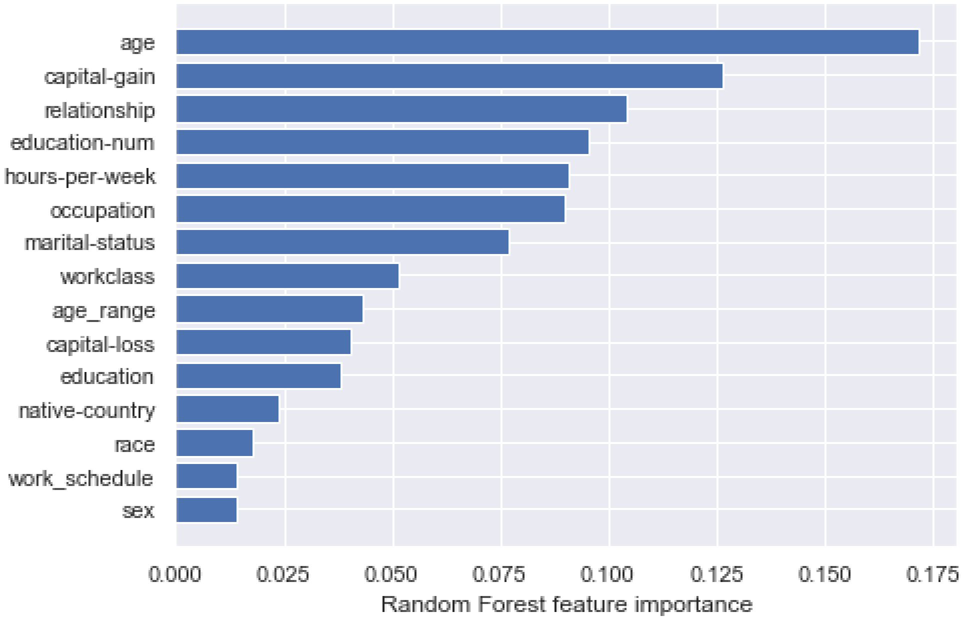


Fig. 14

Again we see many of the same factors we had discussed before as having an influence on income, but the model also puts gender (sex) and education with lower importance. This could be explained because maybe the model is getting most of its information from other variables such as the top ones shown in the chart.