Which factors are the most crucial in determining someone's income? Part-2

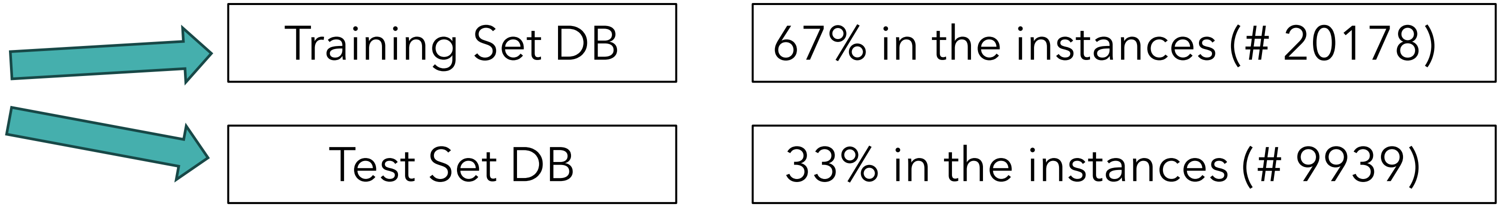
by
Syed Reyadh

Methods

For this analysis I used Scikit learn to perform a decision tree data classification, as I know ex-ante the values of the target variable and many attributes are categorical (including the Target one).

To perform this decision tree the following actions have been taken:

1. I divided the dataset into two:



Training Set DB — 67% in the instances (# 20178)

Test Set DB — 33% in the instances (# 9939)

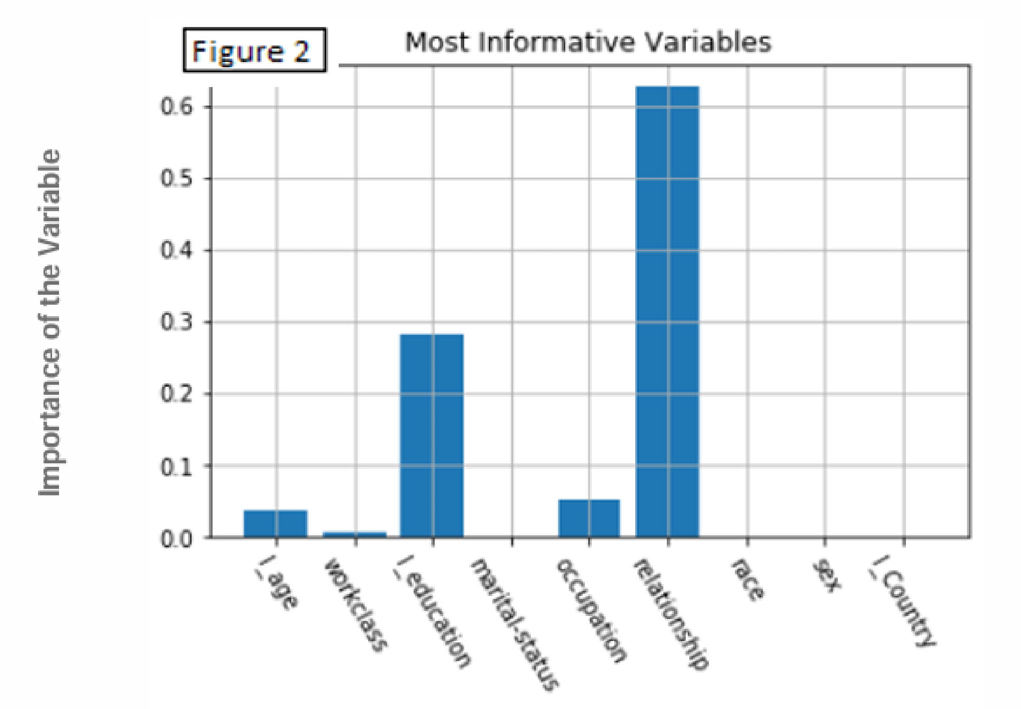
2. I applied the decision tree data classification algorithm to the Training set with the following parameters:

✓ No maximum depth has been imposed

✓ Maximum leaf node == 20

Findings: Results from the Training Phase (1)

After having applied the previously traced model during the training phase, we can determine how much good our attribute are and, in particular, which are the most informative among them, i.e. the attributes that better than other help us in understand the behaviour of our target variable: salary.

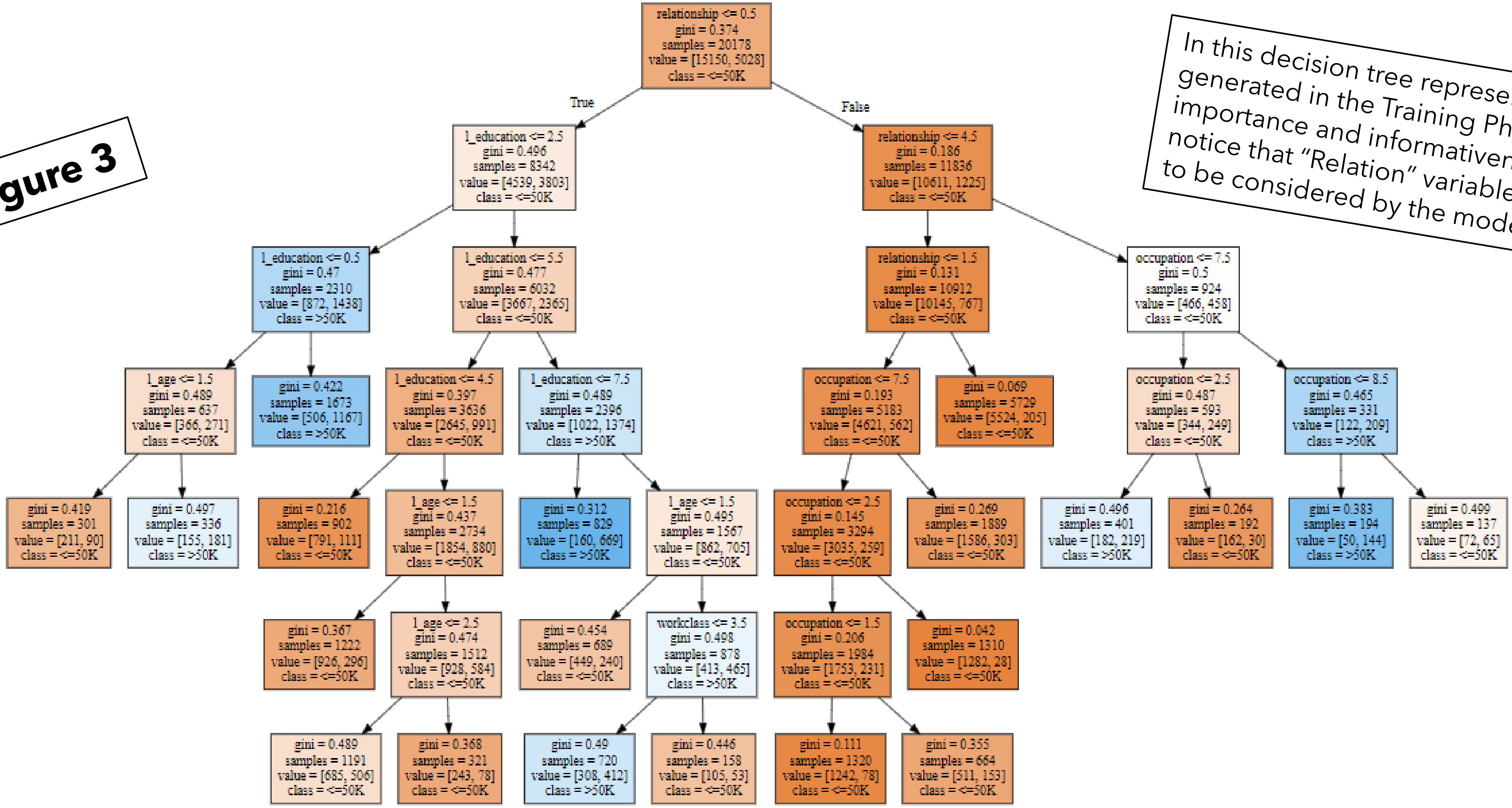


❖In the Figure 2 are displayed the variables that compose our dataset. The information we can get is the level of importance of each of them in terms of informativeness.

- Surprisingly, “relationship” is the most informative variable. It includes observations like “Unmarried”, “Husband”, ”Wife”, “Own-Child”, “not in Family”, “Other Relative”.
- The other most informative variables are Education, Occupation, age, workclass.
- No importance is associated to the other variables: Marital Status, Race, Sex, Country. They seems no to be able to explain income variation across people.

Findings: Results from the Training Phase (2)

Figure 3



In this decision tree represented the rules generated in the Training Phase. Since it's importance and informativeness, we can notice that "Relation" variable is the first one to be considered by the model.

Findings: Results from the Test Phase

✓ The last point remained is: how good the built model is. Is it able to predict the target variable?

✓ To answer this question we can use the test dataset previously generated and try to understand how many mistakes we get by the application of the training phase model.

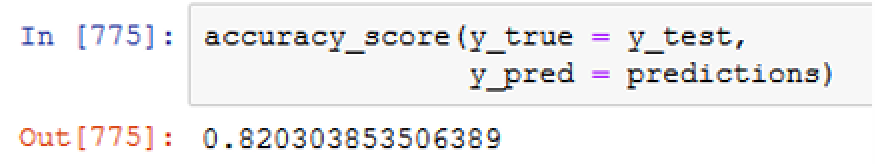
✓ One measure we can use the accuracy of our model is the 'Accuracy Score':

- It is equal to 1 if the model is perfect;
- It is equal to 0 if the model is totally mistaken and no variable is classified correctly;

✓ An accuracy score of 0.75 could be considered good.

Figure 4

Measure of accuracy



```
In [775]: accuracy_score(y_true = y_test,
                         y_pred = predictions)

Out[775]: 0.820303853506389
```

In the Figure 4 is displayed the accuracy score: it is 0.82. It's very good: It's means that more than 82% of the instances that make up the Test set are classified correctly.

Limitations:

Some pertinent limitations apply to this project's work and its outcomes:

1. It’s based on a census done in a town in the USA: it implies that probably it could not be applicable to other different places or Countries.
2. I'm not developing a causality model. This model allows us to conclude that, although there is a correlation, a higher level of education is linked to a greater likelihood of earning a salary of above $50,000. I can't say that earning more money is a direct result of schooling. That is, of course, true in my opinion, but this is not the tool to support it. Richest families, for instance, can afford to provide their kids superior educations. Here, it would be the parents' financial resources—rather than education—that would lead to high earnings.
3. Since the target variable is categorical, the dataset is divided into two, based on the values of the target variable which are either >50k$ or <50k$. This, near the threshold value, generates identification problems for individuals: a person with an income of 49.5k$ should not be too much different from another with an income of 51.5k$. But the model, as it is built, divides them totally. Unfortunately, not having the exact data of income, using a classification model is the only method to derive some information.

Conclusions

- Using Scikit-learn, I created a decision tree classification model that predict the relation between individual characteristics and Salary.
- The level of accuracy of this model is roughly 82%: it means that, knowing ex-ante some individual characteristics, we can predict whether a person will have an income higher than 50K$ or not.
- Among the others, the most important variables that explain differences in salaries are Relationship, Education, Occupation and Age. Remain variables like race, sex, native country seem to be no useful in this.

Acknowledgement

- Data used come from the repository called UCI Machine Learning repository, suggested by the professor Leo during the week 9 of the course of Python for Data Science.

- All the analysis done in this final project come from my effort.

References

- https://www.python.org/doc/
- https://pandas.pydata.org/docshttps://stackoverflow.com/
- https://medium.com/
- https://scikit-learn.org/
- https://numpy.org/
- https://www.graphviz.org/
- Learning Python, 5th Edition, Mark Lutz