Income prediction in census data

Brief

This US Census dataset contains detailed but anonymized information for approximately 300,000 people. The goal of this exercise is to model the information contained in the last column (label), i.e. which people make more or less than $50,000 / year, from the information contained in the other columns. The exercise here consists of modelling a binary variable.

While this task may appear unrelated to the recruitment industry, it aims to surface your understanding of basic data science concepts and practises. In its openness, we want to allow your creativity and unique approach to the surface. Typically candidates take 2-3 hours to complete this.

Please read through the list of takeaways you would be expected to be able to talk through the debrief stage, and budget your time accordingly. We want you to take your time and fit this around your other commitments, but as a guide most candidates respond within a week.

Data

You can access the data through BigQuery, the BigQuery project ID is **heytempo-public** and the dataset ID is data_science_take_home. This dataset has two tables. Access should have already been granted by a member of the data team via a google account you provided to your recruiter. Please reach out if you have any difficulties accessing the database.

You can access the dataset via the Google Cloud console or through a library such as pandas-gbq. More information on this library can be found on their website

```
import pandas_gbq


query = """
SELECT *
 FROM data_science_take_home.census_income_learn
"""


df = pandas_gbq.read_gbq(
 query,
 project_id="heytempo-public",
)
```

Description of the fields you find in the tables, as well as their meanings and information about the dataset itself can be found in the metadata document attached. It is important to read through this file.

Task

Work with Python/SQL to carry out the following steps:

- Import the learning and test datasets
- Based on the learning dataset
  - Perform EDA and univariate audit of the different columns' content and produce the results in visual / graphic format. This audit could describe the variable distribution, the % of missing values, the extreme values, and so on.
  - Create a model using these variables (you can use whichever variables you want, or even create you own; for example, you could find the ratio or relationship between different variables, the one-hot encoding of "categorical" variables, etc.) to model earning more or less than $50,000 / year. Here, the idea would be for you to test one or two algorithms, such as logistic regression, or a decision tree. Feel free to choose others if you wish.
  - Choose the model that appears to have the highest performance based on a comparison between ground truth (label column) and the model's prediction.
  - Apply your model to the test file and measure it's real performance on it (same method as above).

The goal of this exercise is not to create the best or purest model, but rather to describe the steps you took to accomplish it.

Areas and takeaways you will be asked about during the debrief:

- Explain aspects of the task that you found the most challenging.
- Find insights on the profiles of the people that make more than $50,000 / year. For example, which variables seem to be the most correlated with the target variable?
- How does your approach differ from thinking about productionising such a model? Or what modifications would you consider necessary to make in order to use the model.
- How would you explain what you had done to someone who is not a Data scientist?

Finally, you push your code to a Github private repository you can share with us (user **lievcin** in Github).