Feature Engineering and Classification exercise

This exercise is part of the PagSeguro Data Science (DS) interview process, and this material should not be shared with any other person besides the PagSeguro DS or HR (human resource).

Given the dataset at the following address (see below), accomplish some tasks on feature engineering and then apply some machine learning algorithm for classification purpose of census public data.
http://archive.ics.uci.edu/ml/datasets/Census+Income

Requisite

1. Code using **Python** programming language.
2. Use a **Python** Notebook (Jupyter) like environment to your analysis.
If you have a google account, you can use Google Colab, it is very similar to the Jupyter notebook. See more on [1,2].
3. Save or export your notebook locally to upload in the github [3].
4. Use a private account on github to save and share your solution with us
(**ps-datascience** in the github).

What we are looking for in a candidate

1. It's capacity to self-learn.
2. Motivation to learn-fast something that you do not know or do not have practice.
3. Analytical and logical reasoning. Not just someone that mechanically applies tools.

Time and deliverable

You have 7 days to do this exercise. You should produce a report in which you can plot graphics, show tables, and do whatever is necessary to explain your analysis. It is expected that you **show the python code**, it is a report for the DS team. We are not expecting power-point slides or things like that.

Remember that your goal is to **explain your analysis** as a whole in a story-telling approach, which means that the ML classification itself is just one part of the story.

Commit your notebook in a private github account. **Note that it must be private to be accepted**.

Even if you couldn't finish your analysis the way you'd like, we encourage you to submit it anyway, as we'll take everything you could do in the time frame in consideration.

Exercise

Data: http://archive.ics.uci.edu/ml/datasets/Census+Income
Dataset: Predict whether income exceeds $50K/yr based on census data.
Read more about Attributes in the link above.
Follow the tasks as show below:
1. Select Data: collect it. You can save the dataset on your github repo and then load it from the ***notebook***. See snippet 1.
2. Preprocess Data: Format it, clean it. Why not use Pandas! See snippet 2.
3. Transform Data: Feature Engineer. Create new variables as needed and handle missing values.
4. Do some data understanding, and exploratory analysis.
5. Apply some machine learning algorithms for classification of income. Explain your choice of ML algorithm.

Please, we are not looking for some basic ML classification applications. We are not looking for solutions that do a little effort on feature engineering and just apply some ML algorithm from a toolbox. Take your time on feature engineering, feature explanation (which are the most important features?), dataset analysis, and apply some best practices of ML techniques.

It is important to propose a model with reasonable predictive performance. However, make sure you choose a good formulation since we will primarily evaluate you in the logical thinking, methodology and scientific rigor.

Some help!

```
import io
import requests

train_url = 'https://github.com/zzz/train.csv'

s=requests.get(train_url).content
```

Snippet 1

```
import pandas as pd
train=pd.read_csv(io.BytesIO(s))
```

Snippet 2

You can ask for clarification just one time, please send an email to the HR that you are in touch. Remember again, we will consider all your effort! Don't give up!!

Some useful links

[1] https://colab.research.google.com/
[2] https://heartbeat.fritz.ai/getting-started-with-google-colab-notebooks-117e2bb0c220
[3] https://help.github.com/en/github/managing-files-in-a-repository/adding-a-file-to-a-repository
[4] https://towardsdatascience.com/a-quick-introduction-to-the-pandas-python-library-f1b678f34673

How to share your private project