Technical test - Junior / Confirmed Data Scientist

> This test is designed for applicants to junior and confirmed data scientists within iliad group

Objectives

1. Evaluate analytical skills based on a dummy dataset generated for the purposes of the exercise. This dataset is based on a real life use case
2. Evaluate programming skills for analytical purposes (Python highly recommended), including the ability to produce efficient, robust and readable code
3. Evaluate business acumen as well as the knowledge of telecommunication industry
4. Evaluate communication and presentation skills as well as the ability to identify limitations and next steps

Guidelines and deliverables

The deliverables must be sent **at least 2 hours before the interview** to the interviewer via email. The exact format of the deliverables is purposedly left to the apprecation of the candidates, depending on what suits best their needs

There are 2 deliverables:

1. **Conclusions**: the results of the analyses, either as a deck of slides (PowerPoint, etc.) or a text document (Word, markdown, etc.). Candidates are encouraged to illustrate their findings as much as possible (plots, figures, etc.) as well as to back up all their assertions with data
2. **Code**: the code written to produce the analyses (Python file `.py`, Jupyter notebook, etc.)

Both the conclusions and the code can be the objects of a Q&A during the interview.

Candidates **should not spend more than 3 hours** on this test: it is not intended to be a full-fledged study but rather a support for a conversation on the data, the analytical approach and its potential extensions.

Data description

The `load_data.py` file contains an example of Python script to load the data and display the total number of users. It requires the `yaml` package which can be installed via `pip install pyyaml`.

The `fake_liste_abonnes_revo_130122.yaml` file contains the features, as of January 13th, 2022 of the 20k users who subscribed a landline contract in France in January 2016 or January 2017. The selection of the 20k users was made randomly within the 100k eligible users. The sample was restricted to the "Revolution" offer, which comes in several sub-offers with different price patterns.

The following attributes are available for each user :

- `recruit_year_month` : User's subscription date as year-month
- `cancel_year_month` : User's cancellation date as year-month. If the user has not cancelled as of January 13th 2022, then the filed will be 'N/A'
- `total_bill` : total amount billed to user as of January 13th, 2022
- `duration_month` : total subscription duration in months, if user has cancelled
- `fiber_or_adsl` : whether user has a fiber connection or not
- `has_retention` : whether user got a retention offer after contacting the hotline
- `offer` : user's offer
- `sub_offer` : user's sub-offer
- `acquisition_channel` : channel through which the user subscribed

Initial analyses

Please answer the following questions:

Churn:

1. How many users are there in each cohort initially?
2. How many remain in January 2022 ? How many churned each month ? Plot the share of users remaining as a function of tenure. Any difference between user groups ?
3. Compute the monthly churn rate : number of users churning in month M as a fraction of the number of users remaining at the end of M-1. What do you notice ? Why ?

Bills:

1. What's the average total bill? Any difference between user groups ?
2. How does it evolve with tenure ? Why ?

Complementary analyses

Please investigate at least one of the following questions:

1. Can you craft an algorithm to predict which users are going to churn ? How could this be used for business purposes?
2. What is the impact of retention offers ? Should retention measures be generalized ?

Any complementary analyses will also be appreciated.