**Introduction**

The Data Scientist is a critical position for FanAI. The most successful candidate will be joining a fast moving business, that by necessity has to adapt to opportunities, both from our clients and from technology, as it continues to evolve. As a consequence, this is a pivotal role within the company, and in order to ensure we have the strongest possible person in a position, we’d like you to undertake the exercises below. The results will be evaluated both from the perspective of performance, as well as coding quality, including usage of design patterns and object oriented programming.

**Background**

FanAI data comes from disparate sources and can be incomplete or pre-filtered, which can be a disadvantage to us. The purpose of our system is usually to reduce the search space and enrich the remaining dataset, in which to identify a sales prospect. During engagement with a marketing campaign there will be various steps a prospect can go through that give indicators on what stage they are at on a buying journey. Scores are attributed to contacts and collected for each stage. Once scores breach a defined threshold we know it’s time to engage with the sales prospect directly. One of such scores is social reach and influence, that is which brands and eSports players influence a sales prospect.

**Exercise**

**#1 - data manipulation:**

This is a bit of a pure data manipulation/processing exercise. We have provided a sample of some real world data (TrendData.csv) of a value over time.

1. We would like you to come up with the best solution of segmenting the data by date based on trends in the Value column (an example input and output has been given). Please write some form of code/procedure or script to do this, the method should be easily repeatable for new data. The output should be a csv file with the following columns StartDate, EndDate, StartValue, EndValue. Start and End Date define the bounds of the segment and we ask you to include the Value (value from the input data) at the start and end of the trend for convenience. Note: Example\_input.csv and Example\_output.csv only represent the file structures your code should handle, and they are not related at all.

2. List your assumptions and explain why you chose them.

3. Are there any improvements you would potentially make to your approach / what

other information would be helpful in refining/choosing your approach?

**#2 - data processing:**

The goal of this task is to use past data to predict who is going to buy.

Predicting buying intent is one of the most important applications of data science. It allows us to identify the characteristics of a client’s best customers and therefore find more of them all over the world. It also allows us to identify and focus on those potential customers who haven't bought yet, but have high likelihood of buying. Finally, understanding the characteristics of people who have very low intent to buy can be used to identify pain points with the product and use this information to improve it. Predicting buying intent can have a huge impact, both on the marketing side as well as on the product side.

Company XYZ sells routers to companies. It has a database with information about all sales prospects they got in touch with in 2015. Some of these people became customers (i.e. bought at least one router) and some didn't end up buying. You have to: Build a model which assigns to each prospect in the Prospects table the probability of becoming a customer. We will call this score "intent". Please, briefly explain why you chose a given model and the methodology used. Describe the characteristics of high intent vs low intent prospects. How would you use this information to increase sales? Are there any other variables, not included in the dataset provided, that you'd have liked to use for your model?

We have 2 tables included. The 2 tables are:

**Prospects** – general information about prospects and the company they work for

Columns:

• user\_id: the id of the prospect. Unique by prospect. Can be joined to user\_id in

the other table

• company\_sector: the sector in which the company operates, i.e. public administration, finance and insurance, healthcare, etc.

• company\_size: number of employees

You will have

to build a model to predict if a prospect is going to buy and return the probability of buying.

• DMU\_role: Decision Making Unit role of this employee, i.e. Influencer, Budget Holder, Decision Maker...

• joining\_date: when the prospect joined the company

• country: country where the prospect is based

**Sales** – table with all information about the 2015 sales funnel Columns:

• user\_id: the id of the prospect. Unique by prospect. Can be joined to user\_id in the other table

• date\_first\_contact: first time company XYZ got in touch with that given prospect

• source: how company XYZ firstly connected with that given prospect. That is:

email marketing, organic search, referral etc.

• linkedin\_shared\_connections: number of shared connections on LinkedIn between prospect and company XYZ sale representatives

• online\_activity: a score between 5 and 685 of this prospect’s online activity on groups related to routers on LinkedIn/FB/Twitter. The higher this value, the more active the prospect is on social network discussions about routers

• did\_buy : binary variable to be used as a label in the model. 0 means no sale, 1 means the prospect bought at least 1 router in 2016.

Note: Please include your code/script as well as the stored files in your submission. The submission should be done in Python.

For any questions and clarifications, don’t hesitate to contact Victor Genin ([victor@fanai.io](mailto:victor@fanai.io))

Many Thanks and Good Luck!