**Exercise**

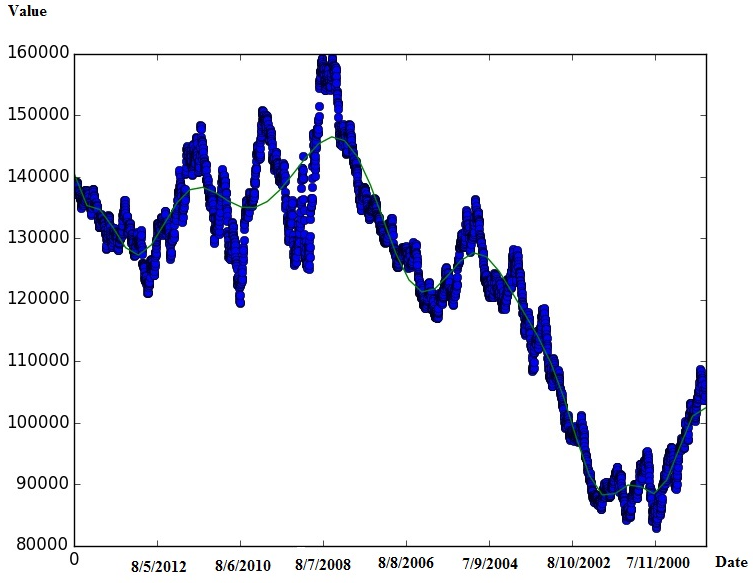
**#1 - data manipulation:**

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.

I used scatterplots to show the trend of data and I fitted a curve on the data to find the minimum and maximum points in the data, where the gradient is zero. Then, I represented the start date, end date, start value, and end value of the data by demonstrating the minimum and maximum of the data.



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

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

This approach works great and shows the trends of the data very well, however, it is not so accurate as it doesn't show all the minimum and maximum points. We could improve it by increasing the degree of polynomial to fit an accurate curve. It is easy to change this point in my code, I just didn’t want to make it more accurate.

**#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.

I used polynomial regression to fit a model for predicting the intent of the customers. In fact, I have twp predictors and one predictand. My predictors are “linkedin\_shared\_connections” and “online\_activity”, “company\_size” and my predictand is did\_buy.

Describe the characteristics of high intent vs low intent prospects.

High intent customers have higher probability of buying a product from this company rather than low intent customers.

How would you use this information to increase sales?

Knowing the customer intent in buying a product helps us to know those customers who are close to leave the company and using churn management and customer management techniques to retain them and also keep track of the ones who are willing to buy the product and be conscious of not loosing them.

Are there any other variables, not included in the dataset provided, that you'd have liked to use for your model?

Yes, there are as many other variables who could help the model to be more accurate, such as age, income, job type, and so on.

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!