Exploratory data exercise

Contents

[Objective 1](#_Toc510517547)

[Business background 2](#_Toc510517548)

[An important technicality 2](#_Toc510517549)

[Your task 3](#_Toc510517550)

[Data provided to you 4](#_Toc510517551)

[Software & modelling approach 5](#_Toc510517552)

[Output and deadline 5](#_Toc510517553)

# Objective

The objective of this exploratory data exercise is to assess:

1. relevant technical skills in the area of machine learning / forecasting and
2. fundamental business understanding.

This exploratory data exercise is not an offer for employment or to contract for the scope included. Issuance of this document, participants’ preparation, participation in meetings and presentation of results and subsequent receipt and evaluation of those results by Novartis does not commit Novartis to award any sort of contract to any participant. Novartis will not be liable for or reimburse any expenses incurred by the participants in relation to the participation in this data exercise. Only the complete execution of a written contract by both parties will obligate Novartis in accordance with the terms and conditions contained in such a contract.

Participants have the right to opt out of the data exercise at any moment without providing any reasons. Novartis reserves the right to select participants and / or exclude participants and / or change the rules and / or the length of this data exercise without providing any reasons.

By participating, participants agree to the rules of the exploratory data exercise.

This document is Novartis Confidential Information which is in addition copyrighted material which may not be copied, recreated or reused in any way or format.

# Business background

Every country organization in the pharma company prepares a forecast for the current month sales it thinks that it will realize. This is done on country level around working day 5 or 6 of each month. This is called “Latest Outlook” (LO).

A concrete example: on March 7 or 8, the country organization in Italy “runs the LO” and forecasts their expected sales for the month of March. The forecast is done in local currency on product – by – product level (“brand level”). The LO is submitted to the global company headquarters (HQ) of the pharma company.

Additionally, every day in the morning, the HQ receives the sales realized by the individual countries on the previous day via an automatic, centralized data feed. This is called “daily sales”. As the month progresses, more and more daily sales become visible and allow to gauge better and better whether the submitted LO is “realistic”, i.e. whether the current months’ “actual sales” (see definition below) will approximately be equal to the LO (say +/-3%), be significantly above or be significantly below.

Once the month is finished, the country organization closes the books for the month and reports the final sales number in the accounting system to HQ. These sales are typically called “actuals” or “actual sales”.

As you can imagine, HQ is interested to know in advance what the actual sales will be. The country LO is a first measure, but as the month progresses, the LO becomes more and more “outdated”. For instance, on March 25, the LO is nearly 20 days old while the daily sales provide already a very good picture of the sales of March with only a few days missing before the month closes.

# An important technicality

Technically speaking, HQ is interested in what is called “3rd party net sales”. These are the effective sales that the company ultimately books, after all rebates, discounts or accounting adjustments (e.g. provisions for product returns) have been deducted. Both actuals and the LO forecast are 3rd party net sales; however, daily sales are not strictly speaking 3rd party net sales and the total of the daily sales may not exactly add up to the actuals of the full month. Read on to understand why.

First, let’s look at a highly simplified value chain in the pharmaceutical industry. A pharma company produces medication and ships/sells it to a wholesaler. For the pharma company, the wholesaler is the “direct customer”. The wholesaler then ships/sells the medication to individual pharmacies. If a patient has received a prescription from a physician, she or he will go to the pharmacy. At the pharmacy, the patient will receive the medication, but it is her or his insurance company (the insurance could be a private or public insurance) that will pay the pharmacy for the medication. This is usually called “the medication is reimbursed”.[[1]](#footnote-1)

Sometimes, the pharma company has a direct agreement with the insurance company. For instance, the agreement could be that the insurance agrees to reimburse the medication to their patients only if the pharma company provides a special discount to the insurance company on this medication. Therefore, the pharma company will receive the sales price from the wholesaler, but at the same time will have to provide a discount to the insurance. The net of these two transactions are the 3rd party net sales.

Daily sales are typically the sales booked with direct customers (wholesales in the example above). Accounting provisions related to discounts as described above (and certain other items) are often booked once a month only, during the closing. Moreover, they may not be booked to direct customers, but to separate provision accounts. Therefore, the sum of the daily sales reported to HQ may not exactly match the actuals of the full month. Note that this is procedure is not “incorrect”, because it is the full month actuals that are ultimately reported in the financial statements. Also, some country organizations have started to book the accounting provisions on a daily basis to minimize this impact on daily sales.[[2]](#footnote-2)

# Your task

Help the HQ to know in advance what the actual sales will be. Based on the data provided (see below), **you need to generate forecasts for the full month actual 3rd party net sales** for the month of:

* November 2017
* December 2017
* January 2018 and
* February 2018.

For each of these 4 months, you need to provide a series of individual full month forecasts. The individual full month forecasts for a given month start on the 15th (inclusive) of the month up to the 25th (inclusive) of the month for every day in that period that is a working day. For each of these forecasts, you can only take the information that would have been available up to this day.

For greater clarity, let’s take November 2017 as an example:

* You need to generate a first forecast for the full month sales of November 2017 on Nov 15 (Nov 15 is a working day). For this forecast as of Nov15, you cannot use the daily sales value from Nov 16 or later, as they would obviously not have been available as of Nov 15 (you can, however, use the LO Nov 17, as this data point would have been available already).
* As a next step, generate a forecast as of Nov 16 using the data up to Nov 16 (ignoring the data from Nov 17 onwards).
* You continue up to Nov 22, which is the last working day up to Nov 25th. There is no separate forecast required for Nov 23-Nov 25, as these are not working days and you don’t have any new information during these 3 days.
* Thus, you create a series of individual full month forecasts for November 2017 which start on Nov 15 and end on Nov 22. These forecasts will probably vary as new information (i.e. daily sales values) flows in over this period. As a result, you should thus have 6 individual forecasts for Nov 2017.

**You also need to generate for each forecast a “measure” which indicates the confidence of this forecast.** This confidence measures indicates how “sure” you are about this forecast. The measure can be a strict, granular numerical value (e.g. probability of being within say +/- x% of actual sales at month end), a more general indicator (e.g. low-medium-high confidence) or a classical confidence interval. You are free to choose the approach that you find most suitable and technically feasible.

The quality of your forecasts will be assessed by calculating the relative variance vs the actual:

Variance = (Forecast – Actual) / Actual

The confidence measure influences also the quality of your forecast; for instance, small variances on forecasts that had a high confidence measure score higher than the same variance with a low confidence measure.

# Data provided to you

We provide you with a brand-country combination. For this combination, you receive:

1. Daily sales covering the following period: Jan 2015 to Feb 2018. Please note that for the 4 “forecast months” (Nov 2017 to Feb 2018), the daily sales values are cut off at the latest on the 25th of these months (exact cut-off depends on when the last working day in the period is).
2. Actual full month sales for the following period: Jan 2015 to Oct 2017.  
   Note that this is the variable that you need to forecast as described in the “Your task” section
3. The LO for the following period: Jan 2015 to Feb 2018   
   This is the LO provided by the country itself at the beginning of the month (around working day 5 or 6). As mentioned previously, you can use the LO of a given month as input to forecast the full month actual sales because the LO is available around workday 5 or 6 each month, before you need to generate your first forecast.  
   Note also that you receive the LO also for November 2017 to February 2018, which you can use for your forecasts; however, the actual full month sales may be significantly different from the LO.
4. A calendar that specifies the working days in each month for the given country. “Working days” are days on which sales to direct customers typically occur, i.e. working days exclude week-ends or public holidays. You have also an overview of the number of particular weekdays in a month (e.g. number of Mondays, Tuesdays,…). This could be relevant in forecasting as certain direct customers have preferred shipping days. For instance, a large direct customer may only buy on Tuesdays and Thursdays, which could lead to higher sales in a months that has relatively more Tuesdays and Thursdays. Please note that it is possible that no sales occur on a working day (i.e. no sales to a direct customer on this day).

The above information is contained in 3 Excel files:

* Exercise - Daily Sales - FOR CANDIDATE.xlsx:  
  Daily sales values (point 1 from above)
* Exercise - ACT and LO Monthly - FOR CANDIDATE.xlsx:  
  Actual full month sales and LOs (points 2 and 3 from above)
* Exercise - Working Days calendar - FOR CANDIDATE.xlsx:  
  Working day calendar (point 4 from above)

# Software & modelling approach

**Your forecasting algorithm needs to be programmed in Python or R**. For your approach to solving the task, you are free to choose whatever statistical or machine learning approach you feel appropriate. You are also not limited to a single model, i.e. you can use different models for different months in the year. However, you need to pick the one forecast and corresponding confidence measure which you feel is the most appropriate.

# Output and deadline

**Please return by email to beatriz.olmos\_gozalo@novartis.com your forecasts and confidence measure in the template (Excel spreadsheet) provided to you, without changing the format of the Excel sheet**. You can describe your approach or additional clarifications and observations in a separate document of your choice (e.g. in the email body, in a separate .doc(x) or .ppt(x) document,…).

**Thank you for participating in the data exercise. We wish you best of luck!**

1. If the process works differently in your country, keep in mind that this is a stylized example to illustrate a specific point about the daily sales. [↑](#footnote-ref-1)
2. Note also that daily sales are generally positive but can also be negative. Negative sales can occur if a customer returns products (and the previously booked sales have to be reversed) or if an accounting adjustment (e.g. for discounts or rebates) is booked. [↑](#footnote-ref-2)