Challenge and Submission Instructions

Table of Contents

1. Sales Forecasting

At Novartis we provide forecasts of the volume that will be sold of each of the brands that we have in our portfolio, and we do that for each country. Each drug in each country has a unique identifier, called cluster id. Each cluster id is independent from one another, even if it might be the same drug being sold in different countries.

2. Lifecycle of a drug

There are multiple crucial moments since the inception of the molecule until the moment the molecule can be released into the market as a product, and multiple others from the moment it is released until the patent runs out and the molecule becomes public, allowing other pharmaceutical companies to produce it, market it, and distribute it as they please.

There are two main phases:

- R&D
- Commercialization

2.1. R&D phase

During the R&D phase, everything is set up so that the product can launch into the market. This year's challenge does **not** focus on the R&D phase.

2.2. Commercialization phase

Overall, we are interested in accurately forecasting the sales at cluster id level for each of the brands in our portfolio. As Eric explained in the video, the product initially grows as it hits the market and it grows up until a certain point. This first phase is known as Growth. Once this growth stabilizes, we reach maturity: this is the period in which a brand is stable in terms of sales: they flatten and look similar month to month for a period of time. At some point there is a Loss of Exclusivity, meaning that the patent runs out and other producers enter the market with the ability of selling the same molecule at a lower price. When this happens, the product enters the decline phase.

During the entire lifecycle, there are also competitors that sell and distribute different products that might serve the same purpose as the Novartis product, and that can also have an impact on the sales of the Novartis drug.

3. The Challenge

3.1. Technical Challenge

The technical challenge is split into two parts, each of them evaluated separately although a strong performance in each of them is required in order to get to the final of the Datathon.

Although, as we have already seen, there may be more than one transition moment, for this challenge we will focus only on the first transition from growth to maturity.

Part 1: Predict transition moment

Predict if there will be a transition to maturity happening for each of the brands during the defined time period, which is in the upcoming 37 months. In case there is, it is important to provide a prediction as to when it will happen.

Part 2: Predict stabilization upon transition

Provide the sales forecast for the same forecasting horizon. Adjust the benchmark forecast, which will be provided to you, based on the expected transition point in time from the first part.

3.2. Business Challenge

Everything that is done from a technical standpoint is backed by a business need, and the end users of our products is always a business group. That means that the results need to be explainable and understandable both to technical and non-technical people, and it is important that the technical solution is consistent but also that the business users for who its built feel comfortable using it and acting on the information that it provides.

4. The Data

4.1. Candidates

- **country:** Country in which the drug is being sold.
- **cluster:** Brand of the drug being sold.
- **cluster_id:** Unique identifier for the country-brand combination.
- **stage_name:** Identified stages and respective category.
- **stage_name_lag_1:** Previously identified stages and respective category.
- **volume:** Number of monthly units sold per drug in each country.
- **business_unit:** Numeric value that identifies the business area of the drug.
- **therapeutic_area:** Numeric value representing the area of action of each drug.
- **prevalence:** Number of patients being treated with a particular condition.

country	cluster_id	date	cluster	stage_name	stage_name_lag_1	volume	business_unit	ther_area_fact	prevalence
country_0	ID_1	2019-04-01	brand_103	growth	no_stage	29M	TWO	1	-1
country_0	ID_1	2020-08-01	brand_103	maturity	growth	19.3M	TWO	1	1569

4.2. Benchmark

Forecasts created through a simple seasonal trend extrapolation based on the recent stage (or the last 18 months).

- **cluster_id:** Country-brand identifier.
- **date:** Forecasted date.
- **fcst_volume:** Forecast generated for each date and cluster_id.

cluster_id	date	fcst_volume
ID_1	2019-04-01	586323465.1
ID_1	2019-05-01	588338833.8
ID_1	2019-06-01	590354202.4
ID_1	2019-07-01	592369571.1

4.3. Population

Contains information on the historical population of a country.

- **country:** Country in which the drug is being sold.
- **year:** Year the data point refers to.
- **population:** Population for each country and year.

country	year	population
country_42	2022	34.07
country_1	2018	12.3
country_4	2019	10.95
country_38	2002	40.7
country_15	2016	11.42

4.4. Regulatory Designations

Indicates if the FDA has given fast track approval to a certain drug.

- **cluster_id:** Unique identifier for a country-brand combination.
- **reg_designation_hasany:** Indicates if the drug has been given the distinctive given if it's high potential.

cluster_id	reg_designations_hasany
ID_1	0
ID_2	1

ID_3	1

4.5. Launches

Information regarding the launch of a country-brand into the market. Each cluster_id has a unique launch date.

- **cluster_id:** Unique identifier.
- **launch_date:** Date when a brand enters one country market.

cluster_id	launch_date
ID_1	2019-12-01
ID_2	2010-01-01
ID_3	2015-03-01

4.6. Indications

An indication is a medical condition for which a specific medication is used. One medication can have multiple indications.

- **cluster_id:** Unique identifier for a country-brand combination.
- **date**
- **indication_entry:** If a cluster_id acquires a new indication at the specified date.

cluster_id	date	indication_entry
ID_1	2020-12-01	0
ID_1	2021-01-01	1
ID_1	2021-02-01	0
ID_2	2019-11-01	0
ID_2	2019-12-01	0
ID_2	2020-01-01	1

4.7. Competitors (Cx)

A competitor is a released product that shares an indication with a Novartis cluster_id. A unique cluster_id can have multiple competitors.

- **country:** Country in which the competitor exists.

- **cluster_id:** Unique Novartis country-brand combination.
- **competitor_entry_date:** Date in which the competitor enters the market.

country	cluster_id	competitor_entry_date
country_1	ID_1	2019-12-01
country_1	ID_2	2020-01-01
country_2	ID_3	2020-03-01
country_1	ID_2	2019-11-01

4.8. Generics (Gx)

A generic is a released product that shares the same molecule as a Novartis cluster_id. They can only appear in the market once there is a loss of exclusivity, meaning that the patent is over. A unique cluster_id can have multiple generics.

- **country:** Country in which the generic exists.
- **cluster_id:** Unique Novartis country-brand combination.
- **gx_entry_date:** Date in which the generic enters the market.

country	cluster_id	gx_entry_date
country_1	ID_1	2021-03-01
country_1	ID_2	2022-01-01
country_2	ID_3	2020-03-01
country_1	ID_2	2020-11-01

5. The metrics

5.1. Metric 1

The first metric evaluates the classification of a cluster_id in terms of if it will transition or not.

Two stages to it:

- If a brand is predicted to transition but it doesn't (False Positive), maximum penalty of 1.
- If a brand is predicted to not transition but it does (False Negative), maximum penalty of 1.
- If a brand is predicted to not transition and it doesn't (True Negative), no penalty.
    - **IT IS VERY IMPORTANT THAT WHEN A CLUSTER_ID IS PREDICTED TO NOT TRANSITION, THE TRANSITION_DATE IS SET TO BE THE FIRST DATE OF THE TIME SERIES FOR THAT SAME CLUSTER_ID.**

  - Otherwise a positive error could be calculated on a True Negative, when the error would otherwise be 0.
- If a brand is predicted to transition and it does (True Positive), the penalty is calculated the following way:

$$metric_1 = e^{\frac{|T_i - G_i|}{F_i}}$$

Where:

- T_i = Predicted date of transition.
- G_i = Ground truth transition date.
- F_i = Forecasting horizon. 37 months.

The final value of the metric will be calculated as:

$$ScaledMetric_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{metric_1 - e^0}{e^1 - e^0}$$

5.2. Metric 2

The second metric will evaluate the accuracy of the forecast once the transition date is provided. It will be a penalized version of the MAPE.

The traditional MAPE is:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|F_i - A_i|}{A_i}$$

In this case, and as explained in the video, there will be a penalization factor introduced to the MAPE:

$$MAPE_{pen} = \frac{1}{n}\left(\sum_{i=1}^{t_0}\frac{|F_i - A_i| \cdot \lambda}{A_i} + \sum_{i=t_0+1}^{n}\frac{|F_i - A_i|}{A_i}\right)$$

6. Platform

6.1. Submission file rules

For both submissions, please replicate the format that has been provided in the sample_submission_metric1.csv and sample_submission_metric2.csv.

- Named columns
- csv file
- No empty values

For the first submission, those cluster_ids predicted to **not** transition must have 2019-04-01 as the transition date, besides having the "NO" in the is_transition column. Also, make sure that there are at least two different transition dates in the file, otherwise it will return an error when trying to submit.

Also, the format for dates has to be YYYY-MM-DD.

6.2. SUBMISSION DATATHON PLATFORM

Internet browser: Google Chrome

URL: http://84.88.76.50

Credentials:

user: teamX@novartisdatathon **password**: pwdteamX

[where 'X' is your team number]

*Please change your password: Click on "Team X" on the top right side *Profile* > *Change password*

Submission:

Keep in mind that you can only perform 3 uploads every 8 hours, in fixed time windows.

First submission:

1. *Dashboard / Panel > Checkpoint – Metric 1*
2. Click the *Upload* button
3. Choose file you want to upload.

Second submission:

1. *Dashboard / Panel > Checkpoint – Metric 2*
2. At the *My Metric 1 Submissions*, choose which Submission 1 upload you want to link to the Submission 2 you are about to upload.

3. Click the *Upload* button.
4. Choose file you want to upload.

Final submission (from 09:30 to 10:30 Sunday 27, November, Central European Time, UTC +1h)

1. *Team Submissions > Select for final*
    a. This will appear at that moment.
2. Press the send selection button.
3. Select a maximum of 2 submissions and send.
4. You can send new selections until 10:30 (the new one will overwrite the previous one)

6.3. Presentation & Code

This only needs to be done by the Top 5 Teams.

1. Download the ppt template in *NOVARTIS DATATHON > General > presentations > Datathon-2022-ppt-Template.pptx.*
2. Prepare your presentation and save it with the name (where X is your team number): *Data_Novartis_Datathon-Results_Presentation_TeamX.pptx*
3. Prepare your code and save it with the name (where X is your team number): *Data_Novartis_Datathon-Final_code _TeamX*
4. Upload your code and presentation in your private channel *TeamX > Files*

For the final presentation, please use the Teams Background provided in *Documents > General > Datathon-2022-Fons-participants.jpg*