7TH edition

The **financial challenge** of the year

NOVARTISDATATHON

online

Challenge and Submission Instructions

Contents

1. The Launches Problem 4
1.1. Context 4
1.2. Launches 4
1.2.1. Definition 4
1.2.2. Importance 4
2. The Challenge 4
2.1. Technical Challenge 4
2.1.1. Objective 4
2.1.2. Data Provided 4
2.1.3. Evaluation Criteria 5
2.2. Business Challenge 5
2.2.1. Integration of Technical and Business Components 5
2.2.2. User Perspective 5
2.3. Presentation to the Jury 5
2.4. Recommendation 5
3. The Data 5
3.1. Data Overview 5
3.1.1. Identifiers 5
3.1.2. Target 6
3.1.3. Drug-Related Features 6
3.1.4. Country-related and Auxiliary Variables 6
3.2. Additional Information 6
4. The Accuracy Metric 7
4.1. Metric Breakdown 7
4.1.1. Error computation 7
4.1.2. Error computation .... **Error! Bookmark not defined.**
4.1.3. Error computation .... **Error! Bookmark not defined.**
4.1.4. Separation of launches 7
4.2. Metric Rationale 8
5. The Platform 8
5.1. Communication Platform: Microsoft Teams 8
5.2. Submission Platform 9
5.2.1. Access 9
5.2.2. Submission Process 9
5.2.3. Final Submission 9

5.2.4. Finalist’s Responsibilities 9
5.2.5. Jury Presentation 9

1. The Launches Problem

In our organization, accurate predictions of newly launched brands are paramount for our finance team. The focus of this datathon challenge, the “Launches Problem”, mirrors a real-world issue we grapple with regularly. As we introduce this challenge to the data science community, we provide participants with an opportunity to address a problem we, ourselves, are actively managing.

1.1. Context

Forecasting the sales of drugs in their early stages is one of the most difficult, yet essential, challenges faced by different companies in the pharmaceutical world. A key challenge is the fact that these drugs have limited to no historical sales data. Traditionally, these forecasts were created using extensive inputs from experts who analyze everything from epidemiology trends, pricing and dosage variations, and similar past brands. For our finance team, this task requires deep, localized knowledge on the drug and on each country the drug is launched.

1.2. Launches

1.2.1. Definition

A new launch is defined as a drug released in a country for less than 12 months, or is projected to be launched within the forecasting year.

1.2.2. Importance

Understanding the sales of a newly launched drug is critical for a pharmaceutical company due to several key reasons. It helps finance teams across a company to estimate expected revenue from a new drug, and accurately informs their financial projections. This, in turn, allows them to manage budgeting, cash flow management and investment decisions, and enables them to identify and manage risks associated with the uncertainty of new product launches.

This challenge tasks participants with developing an innovative solution to accurately predict and analyse the sales of a launch, considering the different country-related and drug-related features associated with the launch. By doing so, we aim to improve forecasting accuracy in order to better inform our business strategies.

2. The Challenge

Now that we understand the significance of the Launches Problem, let's delve into the specific challenges posed by this datathon.

2.1. Technical Challenge

2.1.1. Objective

Participants are tasked to forecast the monthly sales of recent/new drug launches for the year 2023 (test set). The test set consists of brands that have launched in the last 12 months / brands whose launch happens in 2023. If the launch happens in 2023, only the forecast for the rest of the year is required.

2.1.2. Data Provided

To facilitate this challenge, historical launches data of Novartis and other pharmaceutical companies from 2014 - 2022 is provided (train set). The dataset encompasses hundreds of thousands of observations, capturing sales data and other relevant details across various brands, countries and years. This wealth of data offers a rich source for extracting valuable insights.

2.1.3. Evaluation Criteria

Your predictions will be assessed based on their accuracy and effectiveness. The top 5 performers, as determined by specific metrics (will be later discussed), will advance to the next stage—the business challenge.

2.2. Business Challenge

2.2.1. Integration of Technical and Business Components

This datathon uniquely combines a technical challenge with a strong business component. It's not solely about how you solve the problem but also about why you've chosen a particular approach. Understanding the business implications of your solution is crucial.

2.2.2. User Perspective

Keep in mind that the end user of your forecast will be a business team. As such, clarity and explainability are paramount. Craft your solution with a logical business framework, ensuring that the methodology and results are comprehensible to a non-technical audience.

2.3. Presentation to the Jury

After being selected based on technical performance, participants will present their solutions to a jury with both technical and business backgrounds. Therefore, your solution should effectively address both aspects. The top 3 winners will be determined based on this presentation, emphasizing the need for a solution that aligns with both technical excellence and business acumen.

2.4. Recommendation

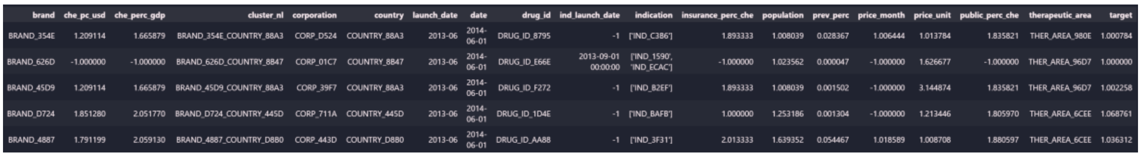
Throughout the process, consider the dual nature of this challenge. Create a solution that not only excels technically but also resonates with the business context. This integrated approach will position you favourably for both stages of the competition and increase your chances of securing a top position in the datathon.

3. The Data

In this section, we will explore the structure and components of the dataset provided for the datathon.

3.1. Data Overview

The data is provided in a single dataframe, and its features can be categorized into four groups: the target, identifiers, drug-related features and country-related features.



brand	che_pc_usd	che_perc_gdp	cluster_nl	corporation	country	launch_date	date	drug_id	ind_launch_date	indication	insurance_perc_che	population	prev_perc	price_month	price_unit	public_perc_che	therapeutic_area	target
BRAND_354E	1.209114	1.665879	BRAND_354E_COUNTRY_88A3	CORP_D524	COUNTRY_88A3	2013-06	2014-06-01	DRUG_ID_8795	-1	['IND_C3B6']	1.893333	1.008039	0.028367	1.006444	1.013784	1.835821	THER_AREA_980E	1.000784
BRAND_626D	-1.000000	-1.000000	BRAND_626D_COUNTRY_8B47	CORP_01C7	COUNTRY_8B47	2013-06	2014-06-01	DRUG_ID_E66E	2013-09-01 00:00:00	['IND_1590', 'IND_ECAC']	-1.000000	1.023562	0.000047	-1.000000	1.626677	-1.000000	THER_AREA_96D7	1.000000
BRAND_45D9	1.209114	1.665879	BRAND_45D9_COUNTRY_88A3	CORP_39F7	COUNTRY_88A3	2013-06	2014-06-01	DRUG_ID_F272	-1	['IND_82EF']	1.893333	1.008039	0.001502	-1.000000	3.144874	1.835821	THER_AREA_96D7	1.002258
BRAND_D724	1.851280	2.051770	BRAND_D724_COUNTRY_445D	CORP_711A	COUNTRY_445D	2013-06	2014-06-01	DRUG_ID_1D4E	-1	['IND_BAFB']	1.000000	1.253186	0.001304	-1.000000	1.213446	1.805970	THER_AREA_6CEE	1.068761
BRAND_4887	1.791199	2.059130	BRAND_4887_COUNTRY_D8B0	CORP_443D	COUNTRY_D8B0	2013-06	2014-06-01	DRUG_ID_AA88	-1	['IND_3F31']	2.013333	1.639352	0.054467	1.018589	1.008708	1.880597	THER_AREA_6CEE	1.036312

3.1.1. Identifiers

- Brand: Identifies the brand associated with the data.
- Country: Represents the country associated with the brand.
- Date: Specifies the date of the recorded data.
- Cluster NL: Country-brand combination ID.

3.1.2. Target

- Target: Monthly sales for the newly launched drug.

3.1.3. Drug-Related Features

- Corporation: The company responsible for the product or treatment.
- Indication: The medical condition the treatment targets.
- Ind_launch_date: The launch date of the indication.
- Therapeutic Area: The field of medicine the treatment relates to.
- Prev_perc: Prevalence, or how common the condition is in the population.
- Price_unit: Price per unit.
- Price_month: Monthly treatment price per patient.

3.1.4. Country-related and Auxiliary Variables

- Population: The total number of people living in the country.
- Public_per_che: The percentage of health expenditure covered by public funds.
- Che_perc_gdp: The percentage of the country's Gross Domestic Product (GDP) spent on healthcare.
- Che_pc_usd: The amount of money spent on healthcare per person, measured in US dollars.

3.2. Additional Information

- As mentioned, data prior to 2023 will be part of the train set, while data from the year 2023 will correspond to the test set. Furthermore, the test set is further subdivided into 2 groups, the public subset and the private subset.
    - The public subset will be used to evaluate all the submission attempts you make, showing the results obtained in each of the attempts. This will give you an idea about your model's performance.
    - The private subset will only be used once to evaluate the final submission you choose. The result of this evaluation will be the final score of your team in this datathon.

  The split between the public and private subsets is randomized at the Country-Brand level with a ratio of 0.5. This means that half of the Country-Brands will belong to the public subset, whereas the other half will belong to the private subset. When we state that a Country-Brand belongs to the public part, we imply that all the months of that Country-Brand will belong to the public part.

- Target Variable: The primary objective is to predict the monthly sales for each launch in the test set.
- Considerations: Exercise caution with potential outliers in the data and prioritize the subset of launches that are weighted more in the metric.

  Understanding the nuances of these features will be crucial for devising effective models to accurately forecast launches sales. Additionally, participants are encouraged to explore the dataset thoroughly, considering the provided hints and incorporating relevant techniques for feature engineering and outlier management.

4. The Accuracy Metric

Understanding the evaluation metric is essential for gauging the performance of your models. The metric used for this datathon is the Composite Year-Month Error (CYME), which is designed to assess performance based on the country-brand for both monthly and yearly levels. This auxiliary function is used in two ways: a weighted CYME for recent launches, and a limited weighted CYME for future launches. The final metric used is the sum of the two weighted CYMEs. Further details are provided in the section below.

$$CYME_L = \frac{1}{2}\left[median_{y\epsilon Y, l\epsilon L}\left(e_{y,l}\right) + median_{m\epsilon M, l\epsilon L}(e_{m,L})\right]$$

$$RLE = w_{RL} \cdot CYME_{RL}$$

$$FLE = w_{FL} \cdot \min(1, CYME_{FL})$$

$$Final\ Metric = RLE + FLE$$

Nomenclature:

$e_{y,l} = yearly\ absolute\ percentage\ error\ of\ launch\ l\ at\ year\ y$

$e_{m,l} = monthly\ absolute\ percentage\ error\ of\ launch\ l\ at\ month\ m$

$Y = set\ of\ years$

$M = set\ of\ months$

$TL = Total\ Launches$

$RL = Recent\ Launches$

$FL = Future\ Launches$

$w_{RL} = \frac{N_{RL}}{N_{TL}} \quad w_{FL} = \frac{N_{FL}}{N_{TL}}$

4.1. Metric Breakdown

4.1.1. Error computation

The average percentage error (APE) between predicted and actual values is computed for all months and years.

4.1.2. Monthly and Yearly Levels

The error is taken for both monthly and yearly level to assess model performance for both granular and yearly aggregated levels.

4.1.3. Median

To account for outliers, the median of the yearly and monthly APE is taken and then combined to compute the CYME.

4.1.4. Separation of launches

A distinction is made between recent launches (brands with less than 12 actuals launched before 2023) and future launches (brands released after 2023). A weight is assigned to each launch depending on the percentage of observations that are recent or future. This weighting system attaches greater significance to the kind of observations that appear more.

Due to the difficulty of predicting future launches, their error is limited to 1. However, participants are encouraged to improve this metric.

4.2. Metric Rationale

The combined CYME error for both recent and future launches provides a holistic assessment of model performance by considering the nuanced importance of two important factors when it comes to forecasting new launches: 1) yearly and monthly accuracy, and 2) whether or not a launch is future or recent.

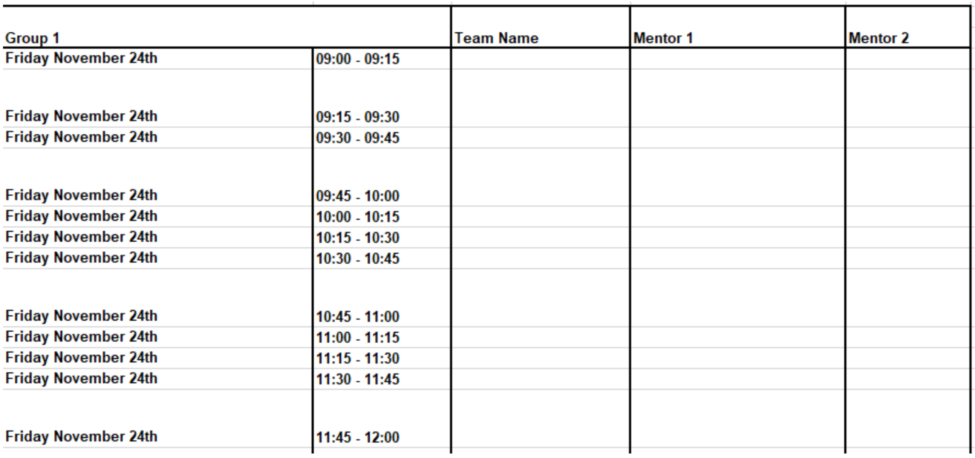
It adapts to the challenge by looking at two crucial levels that a business analyst would consider: the drug's monthly sales phasing, and the yearly total sales. It also acknowledges the extra challenge posed with forecasting a country-brand with no actuals but gives participants the opportunity to go for an interesting solution. Participants should aim for predictions that minimize both yearly and monthly absolute percentage errors for recent launches, and, if time and creativity allow, aim to improve the benchmark metric for the future launches.

5. The Platform

In this section, we will provide details on the communication and submission platform used for the datathon, as explained by Marta and Laura.

5.1. Communication Platform: Microsoft Teams

For effective communication, Microsoft Teams will serve as the primary platform. Two main channels, "Mentoring" and "Novartis Datathon," will be available. The "Mentoring" channel facilitates private communication between teams and mentors, particularly for mentoring meetings. Mentoring meetings will last for 15 minutes, and you will have the chance to book a slot for your team by filling your Team's Name in your preferred slot in an Excel that will be shared on Microsoft Teams (you can see below the structure of this Excel).



Group 1		Team Name	Mentor 1	Mentor 2
Friday November 24th	09:00 - 09:15			
Friday November 24th	09:15 - 09:30			
Friday November 24th	09:30 - 09:45			
Friday November 24th	09:45 - 10:00			
Friday November 24th	10:00 - 10:15			
Friday November 24th	10:15 - 10:30			
Friday November 24th	10:30 - 10:45			
Friday November 24th	10:45 - 11:00			
Friday November 24th	11:00 - 11:15			
Friday November 24th	11:15 - 11:30			
Friday November 24th	11:30 - 11:45			
Friday November 24th	11:45 - 12:00			

On the other hand, the "Novartis Datathon" channel allows open communication among all participants and includes sub-channels for specific purposes. In the "General" sub-channel, only mentors can post, providing general information. The "Mentoring" sub-channel is designed for both mentors and participants to ask and solve general questions. The "Files" tab in this channel includes essential folders, such as "Data" and "Presentations." The "Data" folder contains all necessary files for the competition, including data files (parquet files), metric files for crossvalidation, and instructions on result uploads. The "Presentations" folder includes a template for preparing final presentation slides.

Organizers will be reachable at the times specified in the Agenda.

5.2. Submission Platform

5.2.1. Access

To access the submission platform, participants need to log in with the provided team username and password. The link to the platform can be found in the "submission_instructions" document.

5.2.2. Submission Process

Upon entering the platform, the first step is to change the password for security purposes. This can be done by navigating to the options on the right, clicking "Profile," and changing the password, as demonstrated in the accompanying images. Once the password is changed, teams can begin uploading their submissions. Important: **submissions must be in csv file format**. Please note that teams are allowed a **maximum of 3 submissions every 8 hours**, and it's important to be aware that a failed submission does not count towards this limit. Uploading is accomplished by accessing the "Dashboard/Panel" on the left, clicking on "Checkpoint," and using the designated button for submission. Any error messages indicate issues with the file structure. After the first successful submission, teams will appear in the ranking, and this ranking is updated with each subsequent submission. A "Team Submissions" section allows teams to view and manage different submissions. Given the limited number of attempts to upload submissions, it is crucial to make the most of every opportunity.

5.2.3. Final Submission

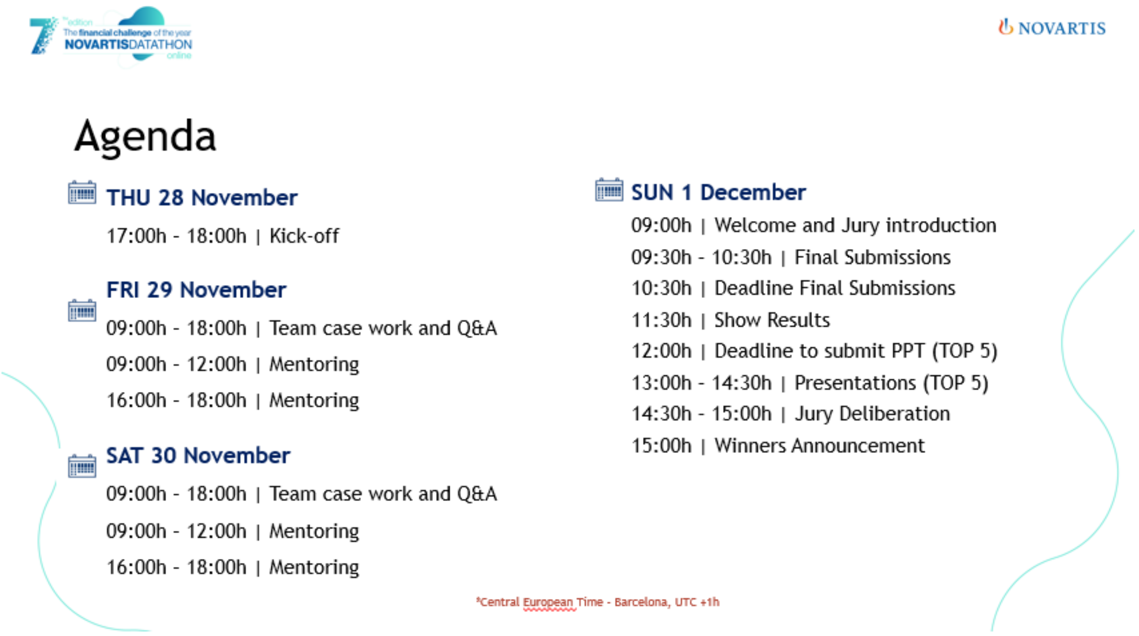
On Sunday at 9:30 a.m., the "Select for final" option will be activated, allowing teams until 10:30 a.m. to choose submissions for final calculation. **The datathon concludes at 10:30 a.m., and no further changes are allowed.** The accuracy metric is then calculated on a private part of the test set. The top 5 results, ordered by the accuracy metric, will be publicized as finalists.

5.2.4. Finalist’s Responsibilities

Finalists are required to prepare a presentation with details on their methodology and results. The platform includes a designated folder for uploading these presentations. Additionally, finalists **must upload the code** used for the final submissions.

5.2.5. Jury Presentation

The presentations will take place between 13:00 and 14:30, and at 15:00, after the deliberation of the jury, the winners will be announced, putting an ending to this 7th Novartis Datathon. Please see the final schedule below



7th edition The financial challenge of the year NOVARTISDATATHON online

NOVARTIS

Agenda

THU 28 November

17:00h - 18:00h | Kick-off

FRI 29 November

09:00h - 18:00h | Team case work and Q&A

09:00h - 12:00h | Mentoring

16:00h - 18:00h | Mentoring

SAT 30 November

09:00h - 18:00h | Team case work and Q&A

09:00h - 12:00h | Mentoring

16:00h - 18:00h | Mentoring

SUN 1 December

09:00h | Welcome and Jury introduction

09:30h - 10:30h | Final Submissions

10:30h | Deadline Final Submissions

11:30h | Show Results

12:00h | Deadline to submit PPT (TOP 5)

13:00h - 14:30h | Presentations (TOP 5)

14:30h - 15:00h | Jury Deliberation

15:00h | Winners Announcement

*Central European Time - Barcelona, UTC +1h

Good luck to you all! 😊