6TH edition
The financial challenge of the year
NOVARTIS DATATHON
online

Challenge and Submission Instructions

Contents

1. The Daily Sales Problem

In our organization, accurate predictions of future sales are paramount for our finance team. The focus of this datathon challenge, the "Daily Sales 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

Our current sales prediction strategy is based on monthly forecasts. However, the dynamic nature of our business requires us to delve deeper into the daily sales patterns to ensure our forecasts align with actual performance. This is particularly crucial when assessing the performance of specific country-brand combinations. The ability to anticipate whether a particular country-brand will overachieve or underachieve against its expected target is a key aspect of our operational strategy.

1.2. Daily Sales Phasing

1.2.1. Definition

Daily Sales Phasing is a vital metric that aids in evaluating the accuracy of our monthly sales forecasts on a day-to-day basis. It is defined as the ratio of sales for each working date within a month. For instance, if a consistent amount is sold every working day, the phasings for that month would average around 1/20, considering a typical work month.

1.2.2. Importance

Understanding the nuances of Daily Sales Phasing is critical because sales distribution is not uniform across all days of a month. The patterns of shipping and sales can vary significantly between countries and brands. For example, while the daily shipping pattern for Country A might predominantly depend on weekdays, Country B's daily shipping pattern could be highly correlated with the working date number.

This challenge tasks participants with developing innovative solutions to accurately predict and analyse Daily Sales Phasing, considering the unique shipping patterns associated with different countries and brands. By doing so, we aim to enhance the precision of our sales forecasts and better align our business strategies with real-time market dynamics.

2. The Challenge

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

2.1. Technical Challenge

2.1.1. Objective

Participants are tasked with predicting the daily phasing for each brand in each country for each of the months of 2022 (test set).

2.1.2. Data Provided

To facilitate this challenge, historical daily phasing data for Novartis brands from 2013 to 2021 is provided (train set). The dataset encompasses millions of observations, capturing daily-level 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

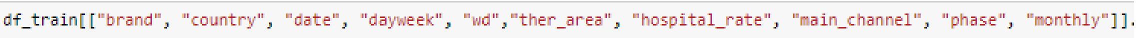
Throughout the process, consider the dual nature of this challenge. Create a solution that not only excels technically but also resonates with the business context. This integrated approach will position you favourably for both stages of the competition and increase your chances of securing a top position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

3. The Data

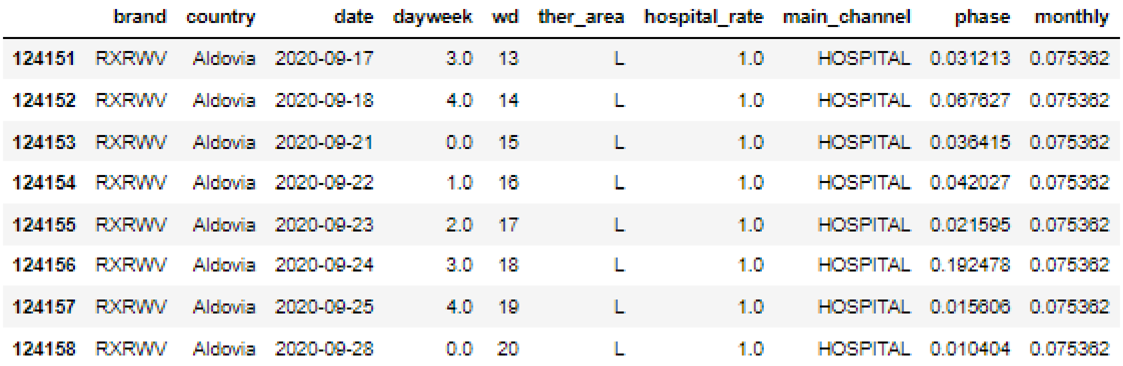
In this section, we will explore the structure and components of the dataset provided for the datathon.

3.1. Data Overview

The dataset is organized into a single dataframe, and its features can be categorized into four groups: identifiers, calendar-related, business-related, and target and auxiliary variables.



```
df_train[["brand", "country", "date", "dayweek", "wd","ther_area", "hospital_rate", "main_channel", "phase", "monthly"]].
```



	brand	country	date	dayweek	wd	ther_area	hospital_rate	main_channel	phase	monthly
124151	RXRWV	Aldovia	2020-09-17	3.0	13	L	1.0	HOSPITAL	0.031213	0.075362
124152	RXRWV	Aldovia	2020-09-18	4.0	14	L	1.0	HOSPITAL	0.067627	0.075362
124153	RXRWV	Aldovia	2020-09-21	0.0	15	L	1.0	HOSPITAL	0.036415	0.075362
124154	RXRWV	Aldovia	2020-09-22	1.0	16	L	1.0	HOSPITAL	0.042027	0.075362
124155	RXRWV	Aldovia	2020-09-23	2.0	17	L	1.0	HOSPITAL	0.021595	0.075362
124156	RXRWV	Aldovia	2020-09-24	3.0	18	L	1.0	HOSPITAL	0.192478	0.075362
124157	RXRWV	Aldovia	2020-09-25	4.0	19	L	1.0	HOSPITAL	0.015606	0.075362
124158	RXRWV	Aldovia	2020-09-28	0.0	20	L	1.0	HOSPITAL	0.010404	0.075362

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.

3.1.2. Calendar-Related Features

- Day of the Week: Indicates the day of the week for a given date.
- Working Date Number: Represents the number of the working date within the month.

3.1.3. Business-Related Features

- Therapeutic Area: Provides information about the therapeutic area to which the country-brand belongs.
- Distribution Channel: Specifies the main distribution channel for the country-brand.
- Percentage of Sales to Hospitals: Represents the percentage of sales made to hospitals.

3.1.4. Target and Auxiliary Variables

- Phasing (Target): The proportion of sales for a specific day relative to the total sales in the month. The **sum of phases** for the same month and country-brand combination should be **equal to one**.
- Monthly Sales (Auxiliary): The sum of sales in a month for a particular country-brand, after being scaled non-linearly. This variable will be used as a weight when evaluating performance. It is **only available in the train set**, not in the test set. You may choose to use it or not.

3.2. Additional Information

- As mentioned before, data prior to 2022 will be part of the train set, while data from the year 2022 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. Giving 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.

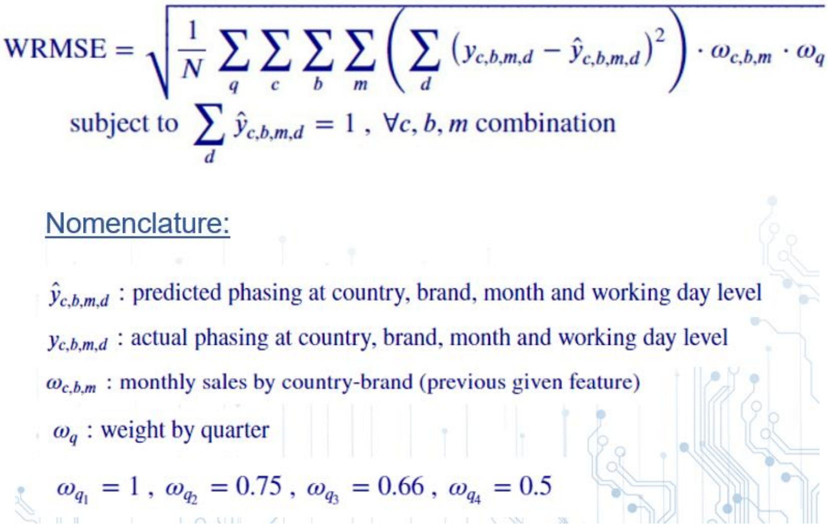
The split between the public and private subsets is randomized at the Country-Brand level with a ratio of 0.4-0.6. This means that 40% of the Country-Brands will belong to the public subset, whereas the remaining 60% of the Country-Brands will belong to the private subset. When we state that a Country-Brand belongs to the public part, we imply that all the months of that Country-Brand will belong to the public part.

- Test Set Considerations: All country-brands in the test set are present in the training set, ensuring continuity in the data.
- Target Variable: The primary objective is to predict the daily phasing (proportion of daily sales) for each brand in the test set.
- Auxiliary Feature: The monthly sales variable, although not available in the test set, will be used as a weight when assessing model performance.
- Considerations: Exercise caution with potential outliers in the data and explore the utilization of lag variables of the phasing for a more comprehensive analysis.

Understanding the nuances of these features will be crucial for devising effective models to predict the daily phasing accurately. 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 Weighted Root Mean Squared Error (WRMSE), which is designed to assess performance at a granular level based on Country-Brand and Month combinations.



$$\mathrm{WRMSE} = \sqrt{\frac{1}{N}\sum_{q}\sum_{c}\sum_{b}\sum_{m}\left(\sum_{d}\left(y_{c,b,m,d} - \hat{y}_{c,b,m,d}\right)^{2}\right)\cdot \omega_{c,b,m}\cdot \omega_{q}}$$

$$\text{subject to } \sum_{d}\hat{y}_{c,b,m,d} = 1\ ,\ \forall c, b, m \text{ combination}$$

Nomenclature:

$\hat{y}_{c,b,m,d}$: predicted phasing at country, brand, month and working day level

$y_{c,b,m,d}$: actual phasing at country, brand, month and working day level

$\omega_{c,b,m}$: monthly sales by country-brand (previous given feature)

ω_{q} : weight by quarter

$\omega_{q_1} = 1\ ,\ \omega_{q_2} = 0.75\ ,\ \omega_{q_3} = 0.66\ ,\ \omega_{q_4} = 0.5$

4.1. Metric Breakdown

4.1.1. Segmentation

The data is segmented based on Country-Brand and Month, allowing for a focused evaluation of the model's performance.

4.1.2. Calculation of Squared Differences

For each combination of Country-Brand and Month, the sum of squared differences between predicted and actual phasing is computed for all working days.

4.1.3. Assignment of Weights

Unique weights are assigned to each Country-Brand-Month combination, directly linked to the sales volume for that specific month within the country-brand. This weighting system prioritizes combinations with higher sales performance, reflecting their greater significance.

4.1.4. Error summation

The errors, calculated for all working days across all months, countries, and brands, are summed up.

4.1.5. Temporal Weighting

Recognizing the diminishing reliability of predictions as time progresses away from the training data (up to December 31st, 2021), a temporal weight is introduced. This weight decreases as the evaluation moves through quarters, adapting to the challenge of limited recent data.

4.1.6. Constraint on Predictions

Since the phasing represents percentages, the predicted phasing must sum to 1 for any country, brand, and month combination. This constraint ensures the coherence and validity of the predictions.

4.2. Metric Rationale

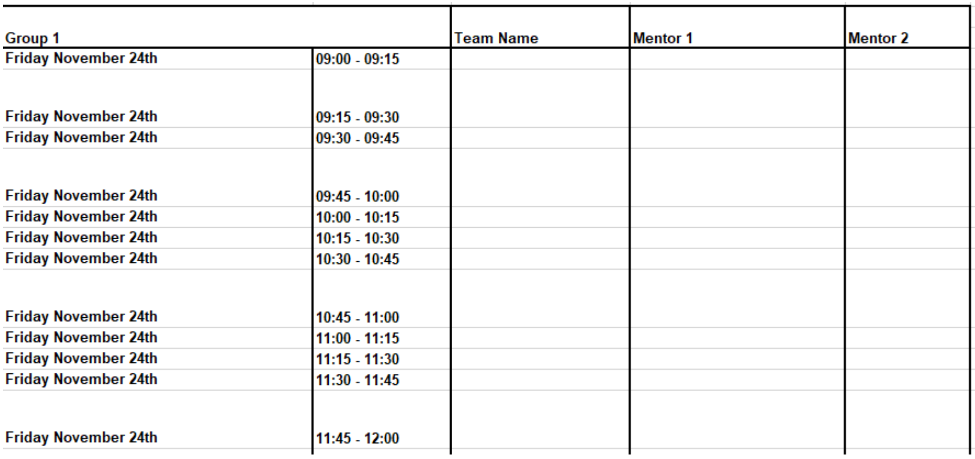
The WRMSE metric provides a comprehensive assessment of model performance by considering the nuanced importance of different Country-Brand-Month combinations. It adapts to the challenge posed by limited recent data in later quarters, offering a holistic view of how well the model aligns with the intricacies of the daily sales problem. Participants should aim for predictions that not only minimize squared differences but also adhere to the constraint of phasing sums totalling 1 for each relevant combination.

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 Mar.

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 cross-validation, 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

Internet Browser: Google Chrome.
URL: http://84.88.76.50
Credentials:
User: teamX@novartisdatathon, password: password to access the kick off
*Please change your password: Click on "Team X" on the top right side Profile > Change password

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

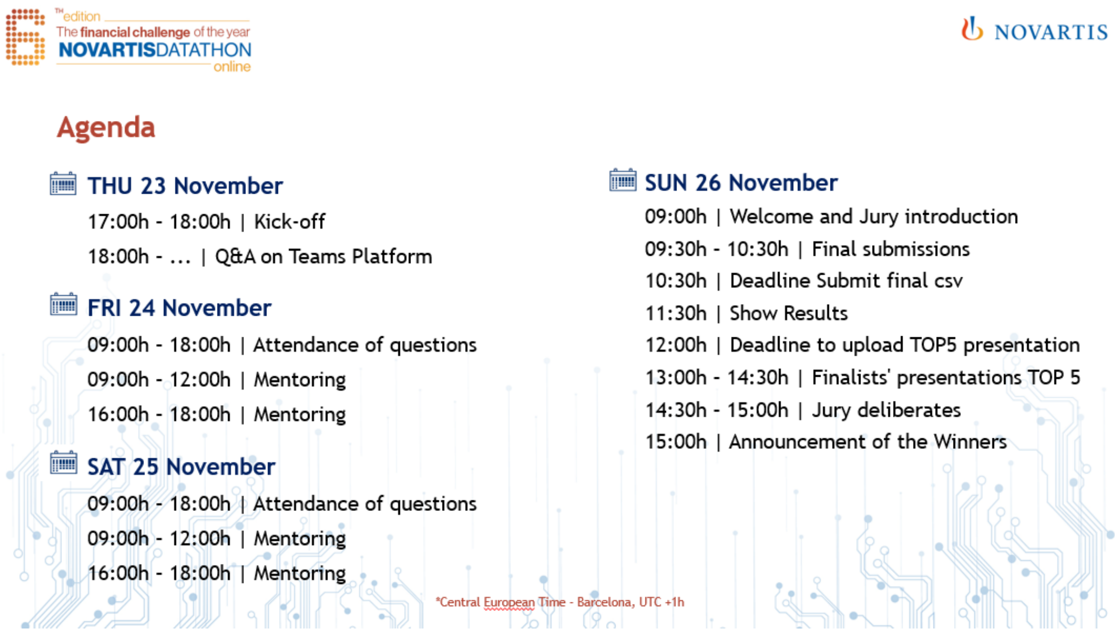
On Sunday at 9:30 a.m., the "Select for final" option will be activated, allowing teams until 10:30 a.m. to choose submissions for final calculation. **The datathon concludes at 10:30 a.m., and no further changes are allowed.** The accuracy metric is then calculated on a private part of the test set. 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 6th Novartis Datathon. Please see the final schedule below



6 th edition
The financial challenge of the year
NOVARTISDATATHON
online
NOVARTIS

Agenda

THU 23 November

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

18:00h - … | Q&A on Teams Platform

FRI 24 November

09:00h - 18:00h | Attendance of questions

09:00h - 12:00h | Mentoring

16:00h - 18:00h | Mentoring

SAT 25 November

09:00h - 18:00h | Attendance of questions

09:00h - 12:00h | Mentoring

16:00h - 18:00h | Mentoring

SUN 26 November

09:00h | Welcome and Jury introduction

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

10:30h | Deadline Submit final csv

11:30h | Show Results

12:00h | Deadline to upload TOP5 presentation

13:00h - 14:30h | Finalists' presentations TOP 5

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

15:00h | Announcement of the Winners

*Central European Time - Barcelona, UTC +1h

Good Luck you all! 😊