

# Data Science Challenge - Predicting User Activity

Nordeus - Job Fair 2024

## About the challenge

A typical Data Science project involves multiple phases, including identifying business opportunities, formalizing the problem, gathering and processing data, developing and training the model, and integrating the solution into a production environment. For this assignment, we have defined the problem and prepared the data, and your task is to develop a model that will use the available data to solve the task and provide relevant predictions.

In this task, we will focus on users who have previously played Top Eleven and decided to return by reinstalling the game. These users, who had played the game before but stopped at some point, are referred to as "re-registrations" when they return to the game. By analyzing their data on the day they re-register (*registration\_data\_training.csv*), along with data from their "previous lives" (*previous\_lives\_training\_data.csv*), we can discover ways to enhance their gaming experience - note that some users may have had more than one previous life.

Information about their previous lives, along with their interactions on the day they re-registered, provides important insights into the behavior of these re-registered users. With this data, we can make more accurate predictions about their future behavior, allowing us to create personalized experiences and content tailored specifically to them. This type of analysis is important for identifying users who have shown they can be active, which helps us understand their needs and expectations better.

## Data description

There are four datasets:

- *previous\_lives\_training\_data.csv*
- *registration\_data\_training.csv*
- *previous\_lives\_test\_data.csv*
- *registration\_data\_test.csv* (similar as training dataset, but without target variable)

You can find the variable descriptions below. Some variables are present in both datasets (*registration data* and *previous lives* dataset):

Variable	Definition
user_id	Unique user number.
registration_time_utc	Date and time of registration.
registration_platform_specific	Platform through which the user had their first session (registration).
registration_country	Country of origin of the first session (registration).

registration_store	The store through which the user had their first session (registration).
registration_season_day	Day in the season when the user registered - one season lasts 28 days.
registration_channel_detailed	Indicates if the user installed the game through T11 marketing campaigns or organically.
registration_device_type	The type of device on which the user had their first session (registration).
registration_device_manufacturer	The manufacturer of the device on which the user registered.
session_count	Number of sessions on the first day of registration.
playtime	Total playing time on the first day of registration (milliseconds).
number_of_devices_used	The number of devices used by the user on the first day of registration.
total_match_played_count	The total number of matches played on the first day of registration.
total_match_won_count	The total number of matches won on the first day of registration.
total_match_watched_count	The total number of matches watched on the first day of registration.
transaction_count_iap	The total number of in-app purchase transactions on the first day of registration.
transaction_count_rewarded_video	The total number of rewarded video ads watched in the game on the first day of registration.
tokens_spent	The total number of tokens spent on the first day of registration.
tokens_stash	The amount of tokens in the stash at the end of the first day of registration.
tokens_bought	The total number of tokens purchased on the first day of registration.
rests_stash	The amount of rests in the stash at the end of the first day of registration.
rests_spent	The total number of rests spent on the first day of registration.
treatments_spent	The total number of treatments spent on the first day of registration.
morale_spent	The total number of morale spent on the first day of registration.
money_stash	The amount of money in the stash at the end of the first day of registration.
avg_stars_top_11_players	The average number of stars of the top 11 players on the first day of registration. Stars represent the quality of players.
avg_stars_top_3_players	The average number of stars of the top 3 players on the first day of registration
avg_age_top_11_players	The average number of years of the top 11 players on the first day of registration.
training_count	The total number of training sessions on the first day of registration.
is_payer_lifetime	Indicates whether the user has made any in-app purchases during their entire lifetime in the game.

days_active_lifetime	The total number of days the user has been active in the game.
transaction_count_iap_lifetime	The total number of in-app purchase transactions in the user's lifetime in the game.
is_rewarded_video_watcher_lifetime	Indicates whether the user has ever watched rewarded videos during their lifetime in the game
days_active_first_28_days_after_registration	The number of days the user has been active in the first 28 days after registration date.

## Target variable

Target variable, *days\_active\_first\_28\_days\_after\_registration*, is an integer ranging from 0 to 28. It represents the number of days a user was active during the first 28 days after re-registration date. A value 0 indicates that the user didn't log in at all during this period, while a value of 28 indicates that the user was active every day within the 28-day period after re-registration.

## EVALUATION

### What is your goal? 🎯

It is your job to predict the number of days a user will be active in the first 28 days after re-registration for each user (identified by *user\_id*) in the test datasets.

### Metric 📊

Submissions are evaluated based on the [Mean Absolute Error \(MAE\)](#) between the predicted number of days a user will be active in the first 28 days after re-registration and the actual number of days the user was active.

### Submission format 📄

You should save your results in the “days\_active\_first\_28\_days\_after\_registration\_predictions.csv” file.

The submission file should contain a row for each user in the test dataset and 2 columns: *user\_id* and *predicted\_days\_active\_first\_28\_days\_after\_registration*. Here is an example:

```
user_id, predicted_days_active_first_28_days_after_registration
1352234, 5
1900923, 28
```

- The submission should be sent via email to [jobfair@nordeus.com](mailto:jobfair@nordeus.com) with a link to your **GitHub** repository (email subject: **Data Science challenge**). Please add your full name to the email! 😊
- Besides the file with predictions, the repository should contain all scripts/ notebooks/ visualizations/images with code that shows how predictions and exploratory data analysis were made.
- For this challenge, you can use the language of your choice, preferably Python or R.

**The challenge is open until November 17, 2024**, end of the day. Good luck!