## **Data Science challenge**

## **Nordeus - Job Fair 2022**

The usual phases tackled while doing ML (Machine Learning) projects are detecting business opportunities, formalizing the problems, getting and cleaning the data, developing and training the model, and integrating all into the production pipeline. For this assignment, we defined the problem, prepared the data for you, and would like to see you develop and train an ML model!

In order to make sure that the T11 (Top Eleven) is a healthy game, we are monitoring many KPIs (Key Performance Indicators) on a daily basis (like DailyActiveUsers, Revenue, Registrations, …). Inspiration for this challenge comes from the fact that while monitoring KPIs we need to have accurate forecasts, before knowing if everything is as expected.

We will provide T11 data describing new users that install the game and the goal is to predict what percentage of them will return to the game tomorrow. This KPI is officially known as Retention Day 1.

### **Data description**

You can find the dataset: [here](https://drive.google.com/file/d/1NZJ1TVBd8TRZJunD7jXmtLZo7JoqeP0t/view?usp=share_link)

| **Variable** | **Definition** | **Key** |
| --- | --- | --- |
| registration\_type | Type of registration:  - Fresh: Users that didn't play T11 before  - Inactive: Users that played T11 before, but didn't progress much  - Active: Users that played T11 before, and had good progress  *\*There is a way to know if user played T11 before (for example, if a user connects social media account)* | 1 = Fresh Registration  2 = Inactive ReRegistration 3 = Active ReRegistration |
| played\_t11\_before | If Fresh Registration, we send in-game popup asking user has he played T11 before.  For other registration types, we do not ask this question. | 0 = data not provided  1 = user says he played T11 before  2 = user says he didn't play T11 before |
| registration\_channel | Label that marks if the new user was targeted by any T11 marketing campaigns, before installing the game.  Users that were not targeted by campaigns, are called “organic registrations”, otherwise, they are called “paid registrations” | 0 = data not provided  1 = Organic  2 = Paid |
| network\_type | Network type used by device, during registration | 0 = data not provided  1 = WIFI  2 = LTE (4G)  3 = HSPAP (3G) |
| device\_tier | Our internal label for the quality of the device used for playing the game | 0 = data not provided  1 = LowestEnd  2 = LowEnd  3 = MidEnd  4 = HighEnd |
| device\_type | Device type | 1 = Phone  2 = Tablet |
| device\_manufacturer | Device manufacturer | 0 = Other + data not provided  1 = Samsung  2 = Xiaomi  3 = OPPO  4 = Huawei  5 = Motorola  6 = Vivo  7 = Realme  8 = INFINIX MOBILITY LIMITED  9 = LGE  10 = TECNO MOBILE LIMITED  11 = Lenovo  12 = OnePlus  13 = TCL |
| screen\_dpi | Device [DPI](https://monopricesupport.kayako.com/article/162-what-is-dpi-and-why-is-it-important#:~:text=DPI%20stands%20for%20dots%20per,screen%20is%20being%20viewed%20from.) |  |
| device\_memory\_size\_mb | Device memory size in MB |  |
| device\_model | Device model name |  |
| os\_version | Device operating system |  |
| registrations | Number of new users registreted |  |
| returned | Number of new users returned to the game tomorrow |  |

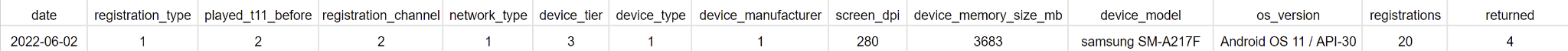
#### Target variable

The target variable is “returned”. It's an integer describing how many users returned one day after the registration.

Once you have the number of registrations and the number of returned, you can calculate Retention Day 1 as

*\*retention\_day1 column does not exist in the dataset*

#### Dataset row explanation:



Out of 20 registrations who came into the game on date 2022-06-01, 4 of them returned on date 2022-06-02. Retention D1 is 20%.

Those 20 registrations are Fresh registrations, who said that didn’t play T11 before, and are targeted by marketing campaigns (such registrations are called “Paid”), registered while the device was connected to the WIFI, and so on…

### Evaluation of the solution

Dataset covers the period between 2022-06-01 and 2022-09-15.

The target variable, for the period after 2022-08-31, is filled with zeros, while all other variables are filled in for the rest of the period.

Your goal is to forecast Retention Day 1, aggregated daily, for the period after 2022-08-31.

Ultimately your result will be evaluated by [RMSE](https://en.wikipedia.org/wiki/Root-mean-square_deviation) for the period after 2022-08-31. Moreover, we will evaluate your approach, including feature engineering.

We would like you to include your code for this assignment and push it to your own GitHub repository.

### **Submission format**

You should save your results in “retention\_d1\_predictions.csv” file.   
The file should contain 16 rows and two columns:

1. date, formatted as “yyyy-mm-dd”, and
2. retention\_d1 values should have percentage format, with 4 decimal points (for example: retention value of 20% should look like 20.000)

We expect the submission to be sent via email to [jobfair@nordeus.com](mailto:jobfair@nordeus.com) with a link to your GitHub repository (email subject: Data Science challenge).

**Challenge is open until November 28, 2022** end of day. Good luck!