Nordcloud
an IBM Company

Problem Statement

During this phase of the recruitment process, your task is to **predict** the probability of a user to click on an advertisement by using the provided digital marketing campaign dataset.

Data Description

The dataset consists of 463.291 entries from a 2021 online advertising campaign.

The training set pertains to the time period of 9.04.2021 - 13.04.2021, while the test set only contains the day of 14.04.2021.

The training set consists of the following attributes:

1. ***Session Id*** – refers to session identifier
2. ***DateTime*** – refers to date and time of the entry
3. ***User Id*** – refers to the user identifier
4. ***Product Type*** – refers to the product type for which the advertisement is
5. ***Campaign Id*** – refers to the campaign identifier of the advertisement
6. ***Webpage Id*** - refers to the webpage identifier
7. ***Product Category*** – refers to the category of the product
8. ***Advertisement Size*** – refers to the area of an advertisement posted on a website, measured in pixels.
9. ***User Depth*** - refers to user's duration of exposure to the advertisement during the respective entry

   (3 being the longest, 1 being the shortest and NA being the inability to measure the time spent)
10. ***Internet Browser Id*** – refers to the identifier of the Internet browser type which is used by the user
11. ***Gender*** – refers to the gender of the user
12. ***Age Group*** – refers to the age group of the user
13. ***City Size*** – refers to the size of the city where the user is located
14. ***Device Used*** – refers to the device used by the user (could be Mobile or PC/Laptop)
15. ***Clicked*** – refers to the fact that the user clicked on the advertisement

16. ***Interested in Cars*** - refers to the fact that the user is interested in cars
17. ***Interested in Food*** - refers to the fact that the user is interested in food
18. ***Interested in News*** - refers to the fact that the user is interested in news
19. ***Interested in Technology*** - refers to the fact that the user is interested in technology
20. ***Interested in Medicine*** - refers to the fact that the user is interested in Medicine
21. ***Interested in Politics*** - refers to the fact that the user is interested in Politics
22. ***Interested in Fashion*** - refers to the fact that the user is interested in fashion
23. ***Interested in Astronomy*** - refers to the fact that the user is interested in Astronomy
24. ***Interested in Animals*** - refers to the fact that the user is interested in animals
25. ***Interested in Travel*** - refers to the fact that the user is interested in travelling

---

Scoring and Evaluation

In order to be evaluated for this task, please send your **solution** as a ".py" or a ".ipynb" file **and** also a ".csv" file with two columns: "Session Id" and "Clicked":

- The first column in the ensuing ".csv" file should match the column of the same name in the submission file provided. We will use these identifiers to match your entries to the reference ones. The identifiers should be the same as those used in the original file. Additionally, in the event of a smaller number of rows than in the original file, the score will be 0.
- The second column should be a numeric column consisting of float values between 0 and 1, which represent the probabilities of successful click events. The aforementioned being valued as follows: 1 meaning that the entry will definitely have a click attributed to it, while 0 meaning there is no chance of a click happening.

Further to these, please send a **report** with all the steps that you have taken during the process (from the very beginning). The report could be either the Jupyter Notebook file with comments for each step or a formal report.

NOTE: You will be evaluated on your ability to **predict** click events and also to **explain the rationale behind your methodology**: the methods used and the extracted relevant insights. However, due to the high complexity of the dataset, a 100% accuracy within the results is not expected.