**Smart Lead Scoring Engine**

Can you identify the potential leads for a D2C startup?

**Problem Statement**

A D2C startup develops products using cutting edge technologies like Web 3.0. Over the past few months, the company has started multiple marketing campaigns offline and digital both. As a result, the users have started showing interest in the product on the website. These users with intent to buy product(s) are generally known as leads (Potential Customers).

Leads are captured in 2 ways - Directly and Indirectly.

Direct leads are captured via forms embedded in the website while indirect leads are captured based on certain activity of a user on the platform such as time spent on the website, number of user sessions, etc.

Now, the marketing & sales team wants to identify the leads who are more likely to buy the product so that the sales team can manage their bandwidth efficiently by targeting these potential leads and increase the sales in a shorter span of time.

Now, as a data scientist, your task at hand is to predict the propensity to buy a product based on the user's past activities and user level information.

**About Dataset**

You are provided with the leads data of last year containing both direct and indirect leads. Each lead provides information about their activity on the platform, signup information and campaign information. Based on his past activity on the platform, you need to build the predictive model to classify if the user would buy the product in the next 3 months or not.

**Data Dictionary**

You are provided with 3 files - train.csv, test.csv and sample\_submission.csv

**Training set**

train.csv contains the leads information of last 1 year from Jan 2021 to Dec 2021. And also the target variable indicating if the user will buy the product in next 3 months or not

|  |  |
| --- | --- |
| **Variable** | **Description** |
| id | Unique identifier of a lead |
| created\_at | Date of lead dropped |
| signup\_date | Sign up date of the user on the website |
| campaign\_var (1 and 2) | campaign information of the lead |
| products\_purchased | No. of past products purchased at the time of dropping the lead |
| user\_activity\_var (1 to 12) | Derived activities of the user on the website |
| buy | 0 or 1 indicating if the user will buy the product in next 3 months or not |

**Test set**

test.csv contains the leads information of the current year from Jan 2022 to March 2022. You need to predict if the lead will buy the product in next 3 months or not.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| id | Unique identifier of a lead |
| created\_at | Date of lead dropped |
| signup\_date | Sign up date of the user on the website |
| campaign\_var (1 and 2) | Campaign information of the lead |
| products\_purchased | No. of past products purchased at the time of dropping the lead |
| user\_activity\_var (1 to 12) | Derived activities of the user on the website |

**Submission File Format**

**sample\_submission.csv**contains 2 variables -

|  |  |
| --- | --- |
| **Variable** | **Description** |
| id | Unique Identifier of a lead |
| buy | 0 or 1 indicating if the user will buy the product in next 3 months or not |

**Evaluation metric**

The evaluation metric for this hackathon would be F1 Score of Class 1.

**Public and Private Split**

Test data is further divided into Public (40%) and Private (60%) data.

Your initial responses will be checked and scored on the Public data. The final rankings would be based on your private score which will be published once the competition is over.

**Guidelines for Final Submission**

Please ensure that your final submission includes the following:

1. Solution file containing the predictions for the id in the test set (Format is given in sample\_submission.csv)
2. A zipped file containing code & approach (Note that both code and approach document are mandatory for shortlisting)
   1. Code: Clean code with comments on each part
   2. Approach: Please share your approach to solve the problem (doc/ppt/pdf format). It should cover the following topics:
      1. A brief on the approach used to solve the problem.
      2. Which Data-preprocessing / Feature Engineering ideas really worked? How did you discover them?
      3. What does your final model look like? How did you reach it?