PROBLEM STATEMENT

Congratulations – you have been hired as Chief Data Scientist of MedCamp – a not for profit organization dedicated in making health conditions for working professionals better. MedCamp was started because the founders saw their family suffer due to bad work life balance and neglected health.

MedCamp organizes health camps in several cities with low work life balance. They reach out to working people and ask them to register for these health camps. For those who attend, MedCamp provides them facility to undergo health checks or increase awareness by visiting various stalls (depending on the format of camp).

MedCamp has conducted 65 such events over a period of 4 years and they see a high drop off between “Registration” and Number of people taking tests at the Camps. In last 4 years, they have stored data of ~110,000 registrations they have done.

One of the huge costs in arranging these camps is the amount of inventory you need to carry. If you carry more than required inventory, you incur unnecessarily high costs. On the other hand, if you carry less than required inventory for conducting these medical checks, people end up having bad experience.

**The Process:**

1. MedCamp employees / volunteers reach out to people and drive registrations.
2. During the camp, People who “ShowUp” either undergo the medical tests or visit stalls depending on the format of health camp.

**Other things to note:**

* Since this is a completely voluntary activity for the working professionals, MedCamp usually has little profile information about these people.
* For a few camps, there was hardware failure, so some information about date and time of registration is lost.
* MedCamp runs 3 formats of these camps. The first and second format provides people with an instantaneous health score. The third format provides information about several health issues through various awareness stalls.

**Favorable outcome:**

* For the first 2 formats, a favourable outcome is defined as getting a health\_score, while in the third format it is defined as visiting at least a stall.
* **You need to predict the chances (probability) of having a favourable outcome**.

Data Description

**Train.zip contains the following 6 csvs alongside the data dictionary that contains definitions for each variable**

Health\_Camp\_Detail.csv – File containing Health\_Camp\_Id, Camp\_Start\_Date, Camp\_End\_Date and Category details of each camp.

Train.csv – File containing registration details for all the test camps. This includes Patient\_ID, Health\_Camp\_ID, Registration\_Date and a few anonymized variables as on registration date.

Patient\_Profile.csv – This file contains Patient profile details like Patient\_ID, Online\_Follower, Social media details, Income, Education, Age, First\_Interaction\_Date, City\_Type and Employer\_Category

First\_Health\_Camp\_Attended.csv – This file contains details about people who attended health camp of first format. This includes Donation (amount) & Health\_Score of the person.

Second\_Health\_Camp\_Attended.csv - This file contains details about people who attended health camp of second format. This includes Health\_Score of the person.

Third\_Health\_Camp\_Attended.csv - This file contains details about people who attended health camp of third format. This includes Number\_of\_stall\_visited & Last\_Stall\_Visited\_Number.

**Test Set**

Test.csv – File containing registration details for all the test camps. This includes Patient\_ID, Health\_Camp\_ID, Registration\_Date and a few anonymized variables as on registration date.

**Train / Test split:**

* Camps started on or before 31st March 2006 are considered in Train
* Test data is for all camps conducted on or after 1st April 2006.