Project supervised learning

# Context – Drunk Smurfs

Among all international hotel guests, Smurfs are burdened with the upkeep of a singular reputation: they are (supposedly) the rowdiest bunch one can entertain, and are equally well-known for unbridled spending as for racking up extensive costs in damages to hotel infrastructure, staff, and occasionally also other guests – costs which typically cannot be recovered once the guest has sought out the safety of his (or her) homeland.

It is your job as a data scientist to screen applying Smurfs clients for an exclusive hotel in the Bahamas - yes, it's the kind of hotel you need to apply for!

# The data

At your disposal is a training set containing data about the behavior of 5000 Smurf hotel guests (train\_V2.csv). This data set contains information about the profit the hotel made during their last visit (excluding damages), but also whether they caused damages during their last visit, and for what amount. These outcomes are respectively called 'outcome\_profit', 'outcome\_damage\_inc', and 'outcome\_damage\_amount'. To predict them, you have access to a host of personal information: previous history of profits and damages, use of hotel facilities, socio-demographics and behavioral scores from the staff of other hotels within the hotel chains. A minor description of features is available in dictionary.csv.

You also get information on the 500 applicants for the 2024 season (score.csv). It is your job to return a list of 150 clients that offer an attractive balance between projected profit for the hotel, and anticipated damages.

You will notice the data set contains a large number of oddities. You are expected to think yourself about what is intuitive and acceptable in terms of approach, and to provide some minor reflection on this in your technical report.

# Possible approach

To generate a client list, you can (but don't have to) follow the next steps:

1. prepare the data set
   * briefly survey the data
   * deal with data issues:
     + appropriate handle categorical data
     + treat missing data
     + identify outliers, and choose whether to make your analysis more robust by removing these
2. predict the projected revenue per clients
   * choose an algorithm, and train it in an optimal way
   * score the 500 applicants
3. predict which clients will cause damage
   * choose an algorithm, and train it in an optimal way
   * score the 500 applicants
4. for those that will wreak havoc, predict the amount of damage they will cause
   * choose an algorithm, and train it in an optimal way
   * score the 500 applicants
5. create a measure of the expected value of each applicant, and create an optimal selection of 200 guests

# Deliverables

At the end, you will present :

1. your full code, as a notebook containing code, comments, and output.
2. the final list of selected clients, with their predicted revenue, predicted damage status (yes/no), predicted damage costs, and overall predicted revenue, as csv.

When delivering notebooks, please provide them in both .ipynb and .html format.