Business Case Exercise

Description

An award winning, internationally recognized gym chain aims at targeting Polish market with the establishment of three new premises. The owners bought and collected data samples on 4000 users.

The data provided covers basic socio-demographics, as well as location data, along with interests in the structured *.csv* (e.g. hobbies), as well as unstructured (e.g. interest groups in the additional *.json* file).

The target variable indicates an initial interest in the long-term gym subscription.

Structured data record

Field	Illustrative Value
user_id	0
target	0
name	Bogusław
sex	male
dob	20643
location	Sosnowiec
location_population	204013
location_from	Lublin
location_from_population	339850
occupation	Refuse workers and other elementary workers
hobbies	Leather crafting
daily_commute	24
friends_number	239
relationship_status	Married with kids
education	4
credit_card_type	Mastercard

JSON data record


```
{
"data": [
{
"groups": {
"data": [
{
"group_name": "Tutoring - will teach / looking for a tutor (Sosnowiec)"
"date_joined": "2012-08-28 05:37:09.743372"
}
{
"group_name": "Easy Sewing for Beginners and Amateurs"
"date_joined": "2009-01-06 11:05:33.816343"
}
{
"group_name": "Homebrewing - the ultimate guide group"
"date_joined": "2011-09-19 07:00:28.366822"
}
]
}
"id": "0"
```


Two samples are provided. The *train.csv* and *train.JSON* sample contains the target (dependent variable) and should be used for model building.

Education is an ordinal variable indicating from secondary (1) to the highest (6) education levels.

Daily commute provided is the approximation of the everyday trip in kilometers.

Expected Outputs

Construct a predictive model with the tool of choice that would predict propensity of a user based on the variables provided.

The test.csv does not contain the target variable. The 'target' should be estimated with your propensity probability (or a binary 1/0 flag) and submitted along with the output report as a scored test.csv file in the following format:

user_id,probability_of_one,target
0,0.567,1
1,0.123,0

The output of the exercise should be a scored file and a report consisting, at the minimum, of the following sections:

- Executive Summary
- Input Data and Transformations
- Model Selection and Training
- Model Quality Assessment
- Findings
- Limitations of the Approach

Be creative in the choice of the tools and the model, the form of the report and applied data transformations.

If you haven't done any predictive modeling, feel free to use data engineering tools to produce cross-tables with business insights for propensity prediction.

Feel free to extend the report by any additional relevant sections, model diagnostics and findings that are applicable to the business case.