PREDICTIVE DATA ANALYTICS FOR BUSINESS

Paris – March, 2025

INSTRUCTIONS FOR THE PROJECT

Data download

• Your learning data set is to be downloaded from Ieseg-online topic 4

(2736_ENG: Project Teams Learning Data Sets | IÉSEG MyCourses)

The dataset is named Project_Dataset_X.sav" where X refers to your project group with the following correspondences :

X	Group number
A	1
B	2
C	3
D	4
E	5
F	6
G	7
H	8
I	9
J	10

• Your Deployment Base is common to all teams: "Project_Dataset_Base.sav".

Python can read sav files,

1) If you are working in GoogleColab, you need to do the following :

```
#Install these two packages

!pip install pyreadstat gdown
!pip install pyreadstat
#Import packages (complete the list if necessary)
import gdown
import pyreadstat
#Put the dateset and on the Google drive, share and paste the share link in the next line (leave everything else unchanged)
url='https://drive.google.com/file/d/1DTXt5IkXatzku1nU5KiilqFABsAOXk2T/view?usp=sharing'
url='https://drive.google.com/uc?id=' + url.split('/')[-2]
output = 'file.sav'
gdown.download(url, output, quiet=False)
df, meta = pyreadstat.read_sav('file.sav')
print(df.head())
```

2) If not, you need to use the following code:

```
#Install the following library
```

```
pip install pyreadstat
# Import the library
import pyreadstat
# Load and read the .sav file as df, example:

df, meta = pyreadstat.read_sav('S1 HealthInsurance Loyalty.sav')
```

0. Context & Objectives

“Customer Personality Analysis” is a detailed analysis of a company’s ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviours and concerns of different types of customers.

Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company’s database, a company can analyse which customer segment is most likely to buy the product and then market the product only on that segment.

You are asked to identify customers that are more likely to consume “wine” and “gold” using the database (1250 observations).

1. Specifications

Your database includes individuals’ characteristics and cultural goods consumption (media …). Two variables of interest (your targets for scoring) were extracted from the e-mail survey:

Wine buyer

Gold buyer

Therefore, you should calculate a score predicting each variable {SW, SG}.

Your campaign will then be a **2-scores campaign**: selected targets should maximize both scores at the same time. A good compromise is maximizing their sum or average, provided they have an equivalent variance (otherwise, standardize them before).

2. Detailed instructions & guideline

2.1. Data download

Your learning data set is to be downloaded from Ieseg-online topic 4: the dataset is named Project_Dataset_X.sav” where X is your project group.

2.2. Understand the data

- describe the data distributions: continuous and categorical variables Descriptive Stats
- describe the kind of relation of all the variables with the 2 target variable S
- evaluate their predictive powers – if any - on target variable S
- crosstabulate Y (in column) with all the explanatory variables Xj (in rows)

2.3. Logistic regression (S1)

Use logistic regression to calculate scores for each target variable (SW1 & SG1).

2.4. Model performance

Evaluate performance: precision, recall, f1 score, accuracy of the following models :

Logistic regression

k-nearest neighbors

Naive Bayes

2.5 : Calculate the AUC for each of the above model

2.6. Deployment

- Download your deployment data base from IESEGONLINE: Project_Dataset_Base.sav

- Apply your selected SW and SG scores and calculate the combined scores on your deployment base
- Simulate operational marketing action (20% top resulting score) on this table and calculate the proportion of expected customers in this base.

- Describe the targeted customers (profile, cultural habits …).

3. Delivery

The deliverables of your research will consist in a pdf file. It will include a description of your project, data, scores construction and performances and a simulation of the upcoming direct marketing action.