ML Classification Assignment

Introduction

Cross-selling is a common selling practice in which additional products or services are offered to current customers. In this assignment you will use Machine Learning techniques on client of an insurance company that has suscribed a health insurance in order to predict if they could be interested in a vehicle insurance provided by the same company.

Dataset Description

You will receive two datasets containing a list of clients with their personal and commercial information. There is a total of 102351 records and 12 explanatory variables divided into two datasets.
• model.csv: the dataset contains the information of 102351 client with the respective target variable. You must use this data to create and evaluate your model.
• predictions.csv: the dataset contains the information of 45196 clients without the target variable. You are requested to provide the predictions for this set of records.

The task is formulated as a binary classification. Your grade will be based on the F1-score metric and on the modeling process presented in the report.

Target Class:

The target attribute is binary: 1 – the customer is interested , 0 – the customer is not interested.

Attribute Information:

n	Attribute	Type	Values
1	id	numerical	ID for the customer
2	Age	numerical	Age of the customer
3	gender	categorical	Gender of the customer
4	Driving_Licence	categorical	0 : Customer does not have a driver licence, 1 : Customer already has driver licence
5	Licence_Type	categorical	Driver licence class
6	Region_Code	numerical/ categorical	Unique code for the region of the customer
7	Previously_Insured	categorical	Customer already has Vehicle Insurance
8	Vehicle_Age	categorical	< 1 Year, 1-2 Year, >2 Year
9	Vehicle_Damage	categorical	1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
10	Annual_Premium	numerical	The amount customer needs to pay as premium in the year
11	PolicySalesChannel	numerical/ categorical	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
12	Seniority	numerical	Number of Days, Customer has been associated with the company
13	Target	categorical	1 : Customer is interested, 0 : Customer is not interested

Submission Instructions

1. Model Training Data Release: 12 March 2021, 20:00.

2. Description of analysis on the training set and model identification: 16 March 2021 20:00.
You are asked to kindly send an email with the following supporting information:

a) A **brief report** of the step-by-step methodology (i.e. pre-processing, visualization, training, testing, etc.) that you have followed to develop your model, this document must illustrate the motivation behind your selected approach.

• File Format: .pdf • Filename: surname1_surname2_surname3.pdf (e.g. orsenigo_soto.pdf)

b) **The commented python code** that you used in your model. Comments in the code must ensure that the code is easy to follow.
• File Format: .ipynb, .py • Filename: surname1_surname2_surname3 (e.g. orsenigo_soto.ipynb or orsenigo_soto.py)

3. Prediction Data Release: 19 March 2021 20:30.

4. Prediction Submission: 21 March 2021 20:00.
You are kindly requested to strictly follow the described submission guidelines:
• File Format: .csv
• Filename: surname1_surname2_surname3 (e.g. orsenigo_soto.csv)
• Column Format: **A single** column named "target"
• Row Format: Your predictions (0 or 1) with **the same number of rows** and in the same order as the **prediction** test set.
Example:

id	Gender	Age	Driving_License	Licence_Type	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Seniority
1	Male	51	1	M	28	No	1-2 Year	Yes	45301	26	31
2	Male	63	1	B	39	No	1-2 Year	Yes	18609	124	28
3	Male	76	1	M	38	No	1-2 Year	Yes	35081	26	164
4	Female	38	1	B	41	No	1-2 Year	Yes	19603	124	33
5	Female	21	1	D	30	Yes	< 1 Year	No	27842	160	28
6	Female	30	1	B	28	No	1-2 Year	Yes	26357	157	204
7	Female	27	1	C	49	Yes	< 1 Year	No	17510	152	298
8	Male	25	1	B	8	Yes	< 1 Year	No	27307	160	250
9	Male	42	1	A	28	No	1-2 Year	Yes	37731	124	298
10	Male	43	1	C	36	No	1-2 Year	Yes	32840	124	207
11	Female	23	1	C	47	Yes	< 1 Year	No	37578	152	48
12	Male	38	1	M	28	No	1-2 Year	Yes	37721	26	176
13	Female	22	1	D	37	No	< 1 Year	Yes	29793	152	282
14	Female	41	1	M	28	No	1-2 Year	Yes	35821	124	222
15	Male	27	1	C	8	Yes	< 1 Year	No	53029	152	109
16	Male	72	1	B	28	No	1-2 Year	Yes	85517	124	281
17	Male	30	1	A	29	Yes	< 1 Year	No	36658	152	227
18	Female	26	1	M	50	No	< 1 Year	Yes	31209	154	278
19	Male	25	1	B	8	Yes	< 1 Year	No	2630	151	286
20	Male	25	1	B	29	Yes	< 1 Year	No	32927	152	31
21	Female	48	1	D	46	No	1-2 Year	Yes	2630	26	178
22	Male	25	1	C	29	Yes	< 1 Year	No	37864	152	222
23	Male	27	1	M	6	Yes	< 1 Year	No	19928	152	17
24	Male	46	1	D	28	No	1-2 Year	Yes	59494	26	214
25	Male	46	1	M	28	No	1-2 Year	Yes	33527	26	203
26	Male	70	1	M	46	Yes	1-2 Year	Yes	2630	26	84

target
0
0
0
1
0
0
0
0
1
0
1
0
0
0
0
0
0
0
1
0
0
0
0
1
0
0

Further Instructions

• **Any submission that does not respect the guidelines (submission after deadline, empty file, wrong student code) will not be graded**.