ML Classification Assignment

Introduction

Cross-selling is a common selling practice in which addittional 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 comercial 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: 06 November 2020, 20:00.

2. Description of analysis on the training set and model identification: 14 November 20:00.
You are asked to kindly submit in the Beep page 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: 6-digit student code (e.g. 123456.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: 6-digit student code (e.g. 123456.ipynb or 123456.py)

3. Prediction Data Release: 14 November 21:00.

4. Prediction Submission: 16 November 20:00.
You are kindly requested to strictly follow the described submission guidelines:
- File Format: .csv
- Filename: 6-digit student code (e.g. 123456.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 test set.

Example:

Further Instructions

- The assignment can be developed in groups with a maximum number of three participants. Nevertheless, **submission is individual**, therefore each student must upload his/her own submission files (even if they are the same for all participants in the group).
- **Verify** your uploaded files in the Beep platform.
- **Any submission that does not respect the guidelines (submission after deadline, empty file, wrong student code) will not be graded**.