

AI DATA SCIENTIST CHALLENGE

VPINS is a division in VPB, specializing in providing insurance services with many attractive policies. Currently the company is looking to promote incentive for a special kind of healthcare insurance called AIA policy. In this challenge, you are going to help the division manager to understand their customers by supporting him to identify which customer is willing to possess the policy, so he can make an efficient sales campaign. He expects the final output of your work as follows:

1. A maximum 2-A4-page document (or maximum 10-page slide) presenting your problem understanding, approach, workflow, results.
2. Your code base using any programming language of choice (preferably Python) via a compressed file so other data scientists can replicate your work.

Please contact haont89@vpbank.com.vn for any further questions.

Disclaimer:

The information contained in this document is confidential, privileged and only for the information of the intended recipient and must not be used, published or redistributed without the prior written consent of the EDA-AI Center

DATASET DESCRIPTION

1) *train_data.txt*:

You will use this dataset to train and validate your prediction models. Each row of this file consists of 86 attributes including sociodemographic (attributes 1-42) and product ownership (attributes 43-85), the attribute 86 is the target variable or label (possessing a AIA policy or not) and ID of each customer (identical with the line number of the row). The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same socio-demographic attributes

2) *test_data.txt*:

This dataset is for predictions. It has the same format as *train_data.txt*, except that the target variable is missing.

3) *attributes_description.pdf*:

This file contains the description of each attribute in detail.

TASKS

1) Prediction task: The objective of this task is to construct a model that predicts whether a customer will buy an AIA policy or not. You are also provided a test set (file *test_data.txt* above) of 4000 instances who were likely to have the policy. Please filter out the most 800 highly promising cases that want to buy the mentioned policy. We will use the hold out label set to verify your results, the label is not provided to you, so the more cases that really want to buy an AIA policy you can detect, the better the model you built.

The output of this task is a file that contains 800 lines, each line is the ID of a customer.

2) Explanation task: The objective of this task is to provide insight into why customers have an AIA policy. The explanation and accompanying interpretation should be scored on comprehensibility, usefulness and actionability to effectively support the manager in making his sales campaign