Puzzle Overview

Congratulations on arriving at this stage of Nubank's Data Science Team hiring process! We are super excited to receive and talk about your solutions and ideas!

Let's begin: This is a Kaggle like test and you are supposed to do it as if you were a data scientist here. Our goal is to evaluate how you would solve a problem related to our business and your knowledge of Machine Learning.

Please, follow the instructions below strictly.

Context

As a credit company, it is important to know beforehand who is able to pay their loans and who is not. The goal of this puzzle is to build a statistical/machine learning model to figure out which clients are able to honor their debt.

Structure

Besides these instructions, the ZIP file contains 2 datasets in CSV format:

- puzzle_train_dataset.csv: training dataset
- puzzle_test_dataset.csv: test dataset. It contains the same columns as the training dataset, except for the `default` variable.

Your goal is to predict the probability of default, which is identified by the default variable in the training dataset.

Deliverable

You must send us an email with (keep in mind that all the items below are REQUIRED):

- an explanation of what you did (maximum length of 2000 characters with spaces);
- your own code, which we should be able to run (you can assume that we have access to the same resources that you used). Please include the explanation at the beginning of your code;
- instructions about how to run the code (e.g., software, programming language, required dependencies, commands to run);
- a .csv file called predictions.csv with 2 columns: the IDs from puzzle_test_dataset.zip and the predicted probabilities under the column predictions;
- the expected performance score of your predictions;
- an explanation of how the model's output prediction could be used to make credit decisions.

We strongly recommend you restrict all the code, including comments and analysis to a single notebook/markdown file. You can solve this test in Python or R, and you are allowed to create your own libraries in separated files as you wish.

Don't identify yourself in any files! We do blind reviews so people are not supposed to know who you are while evaluating your work. Be careful with code warnings, since your name could be written in folders!

Please try to follow these instructions as closely as you can. What we will evaluate your ability to:

- Handle, clean, explore and analyze data;
- write good quality code (e.g. reproducible, readable);
- apply machine learning models to real problems;
- clearly communicate your decision process;
- split complex and real problems into solvable pieces.

We hope you enjoy solving this challenge and please let us know if you have questions or want to give us feedback.

Nubank's Data Science Team