réinventons / notre métier

Actuary and Data-Science Study Case

AXA France - Direction Offre IARD Particuliers et Professionnels

Objective

The general principle of this study case is similar to a “Kaggle” competition: the aim is to construct a predictive model from a given dataset.

We propose to estimate the expected annual profit per customer on an automotive insurance contract according to the characteristics of the customer.

Instructions

You have normally received 2 csv files in addition to these instructions:

1. ***Labeled_dataset.csv*** 1 contains variable names online, and 1000 data lines. Each of these 1,000 lines contains:
    - 1 index from 0 to 999, uniquely representing each client;
    - 12 variables (age, salary, bonus-malus coefficient ...) characteristics of the client;
    - 1 "target label", profit, indicating the annual net profit realized for that client;
2. ***Scoring_dataset.csv*** 1 contains variable names online, and 300 data lines.

Each of these lines contains:

- 1 index, from 1000 to 1299;
- 12 variables *(age, salary, bonus-malus coefficient ...)*

You must use the ***labeled_dataset*** data to build a predictive model. You can then apply this model on the ***scoring_dataset*** to generate your predictions.

These predictions will be evaluated by our team as we have the labeled dataset containing the "real" realized profits.

The metric used to compare your results with the answers will score the **RMSE** (Root Mean Squared Error).

IMPORTANT INFORMATION ABOUT THE DATA:

The dataset for this test was entirely generated by our team. We tried to provide a business connotation in connection with insurance, but all data are purely fictitious.

More specifically, ages, wages, socio-professional categories, etc. are not representative of our customers or of any population. The link between the characteristics of a client and the realized profit is also fictitious. Achievable accuracy in terms of performance of the model is not necessarily realistic either.

Technological Restrictions

The choice of technology is left to your discretion.

Here is a non-exhaustive set of relevant languages and libraries: Python, Scikit-learn, R, Apache, Spark, Java, Scala, C ++... It is recommended to use a technology with which you are comfortable.

The quality of the code is an evaluation criterion.

It is imperative that we can execute the code provided. **This rule excludes paid license languages or libraries** (e.g.: SAS, SPSS ...). It is advisable to use open source technologies.

Expected Deliverables

1. **A csv file containing your predictions:** This file must be composed of 300 lines, each containing the index *scoring_dataset.csv* and your prediction for this index.

2. **The code** that enabled the generation of the model and instructions for running the code if necessary.

3. **A presentation** of the **methodology** in place and the **results obtained.** The

Format of this document is imposed: PDF file or MS PowerPoint presentation.

Evaluation criteria

1. The **RMSE score** of your submission, based on our core test.
2. The **methodology** implemented is at least as important as the score itself. It is thus important to explain your approach. In particular, do not hesitate to

mention other methods that you have tested, including those which have not proved to be relevant.

3. **Quality of code:** readable code and comments is highly valued.
4. Last but not least : the **quality of the written restitution** will be decisive.

Additional details

If you have any question, please contact Julien Durand: Julien.durand@axa.fr

Happy data crunching !