Introduction to Finance for Data Scientists

Group Work: Lending Game

Context

You run a fintech lending platform making consumer loans to individuals. You receive loan applications from a large number of individuals and must decide whether and at which interest rate you offer to lend money to each loan applicant. You are in competition with two other lenders (two other teams of students) who also make loan offers to these individuals. Borrowers may default on their loan, so it is important to choose the interest rate at the right level to compensate for the risk of default.

You have information about loan applicants to estimate their risk of default. Some of the information about loan applicants is available to all lenders. Each lender also has its own source of private information about loan applicants extracted from digital footprints. Because different lenders use different algorithms, they obtain different private signals capturing different dimensions of loan applicants' creditworthiness. All lenders also have access to data on past loans to train a credit scoring model.

Data

Past loans. All lenders have access to the same dataset `PastLoans.csv`, which contains information on loans made in the past for which it is known if the borrower eventually defaulted or not. The data include the following variables:

- `sex`: 0= female; 1= male
- `employment`: employment status (four categories)
- `married`: 1= married; 0= other
- `income`: annual income in euro (top coded at 1M euros)
- `digital1`: digital footprints extracted by lender 1 (coded from 0 to 1)
- `digital2`: digital footprints extracted by lender 2 (coded from 0 to 1)
- `digital3`: digital footprints extracted by lender 3 (coded from 0 to 1)
- `default`: 1= the borrower defaulted on the loan; 0= the loan was repaid

Note that past loans data contain the digital footprints variable of all three lenders. The reason might be that the banking regulator mandates lenders to disclose their information after a certain amount of time (this is not the case in practice). Another reason might be that lenders have been hacked and their data leaked in the public domain. No matter what the reason is, the fact is that all lenders have access to the exact same data on past loans to train their credit scoring model.

NB: All these data are simulated data, not real data. However, they are meant to "feel" real in terms of how the different variables are distributed in the population and how they correlate with default.

New loan applications. All lenders receive the same 100,000 new loan applications from new potential clients. Each loan applicant asks for a loan of 10,000 euros. The three lenders do not have the same information set about new applicants. While all lenders have the variables sex, employment, married, and income in their database, each lender only observes its own digital footprints variable. The loan applications are in the file `LoanApplications_xxx.csv`.

The new loan applicants are drawn from the exact same population as the borrowers in the first dataset. In particular, the determinants of default are exactly the same in the first dataset and in the new pool of loan applications.

Organization of the Loan Market

Interest rate. Your job is to decide whether and at which rate you make a loan offer to each of these 100,000 loan applications. An offer is an interest rate at which you would be willing to make a loan.

Loan applicants select from which lender they take a loan as follows. Denote the three lenders by $k = 1, 2, 3$ and the loan applicants by $i = 1, \ldots, 100000$. Denote the interest rate offered by lender k to loan applicant i by $r(k, i)$. For example, $r(k, i) = 0.05$ if the interest rate is 5%. If the lender makes no offer to this applicant, we denote $r(k, i) = \infty$. There are three types of loan applicants:

- Type 1 (one-third of the loan applicants): They have a preference for lender 1 and are ready to pay 2% extra to get a loan from lender 1. Formally, they choose the lowest among $r(1, i) - 0.02$; $r(2, i)$; and $r(3, i)$.
- Type 2 (one-third of the loan applicants): They have a preference for lender 2 and are ready to pay 2% extra to get a loan from lender 2. Formally, they choose the lowest among $r(1, i)$; $r(2, i) - 0.02$; and $r(3, i)$.
- Type 3 (one-third of the loan applicants): They have a preference for lender 3 and are ready to pay 2% extra to get a loan from lender 3. Formally, they choose the lowest among $r(1, i)$; $r(2, i)$; and $r(3, i) - 0.02$.

For example, if lender 1 makes a loan offer at 5%, lender 2 at 6%, and lender 3 at 8%:

- A type 1 applicant takes lender 1's offer at 5%.
- A type 2 applicant takes lender 2's offer at 6%.
- A type 3 applicant takes lender 1's offer at 5%.

Therefore, a loan offer you make is not necessarily accepted. It can be rejected because one of your competitors makes a cheaper offer to the same borrower or because the borrower has a preference for another lender. Conversely, some borrowers may accept your offer even if it is not the cheapest.

Payoffs. When a borrower chooses your loan offer, your profit on that loan depends on the interest rate you offered and on whether the borrower defaults or not. Therefore:

- If the borrower does not default, you earn the interest rate you offered times the size of the loan. Your profit on that loan is $r(k, i)$ times 10,000 euros.
- If the borrower defaults, you lose the amount you lent (the recovery rate is zero). Your profit on the loan is negative 10,000 euros.

Your total profit is the sum of the profits and losses you make on all the borrowers who take your offer.

Instructions

Stage 0. Please email the composition of your team of 4±1 students before the morning's class on October 10 to hombert@hec.fr.

The game and assignment takes place in two stages.

Stage 1. In the first stage, your job is to predict the probability of default of the loan applicants and to decide whether and at which rate to make them offers. You must also choose a fun name for your fintech (marketing matters too!)

Please submit:

1. The name of your fintech.
2. A csv file with the list of the 100,000 loan applications and two columns containing the variables
    - `id`: loan applicant identifier provided in the data set `LoanApplications_xxx.csv` (running from 1 to 100,000)

- `rate`: interest rate you offer to the applicant. Please input 0.12 for an interest rate of 12%. You are not allowed to offer interest rates above 100%. If you don't want to make an offer to a loan applicant, leave the interest rate variable missing for this applicant.

Your input for the first stage is due on **October 21** at 24:00 by email to hombert@hec.fr.

Stage 2. At the end of stage 1, I will use the loan offers made by your team and the other teams to simulate the outcome of the market: which lender does each applicant choose, whether a default occurs or not, and the total profits made by each team. I will send the results to each team, along with the complete dataset. This information will allow you to figure out whether you made money, and why, or why not.

In stage 2, you receive another 100,000 new loan applications and play the lending game a second time. Of course, you should learn from the experience of the first stage and try to improve your strategy. You are asked to:

1. Submit a csv file with the list of the 100,000 new loan applications, again with the interest rate you offer to each applicant.

2. Submit a 3-page report (an actual text, not slides or bullet points) explaining:

   a. Your methodology for estimating the default probability and how you chose the interest rate in stage 1. In particular, explain the problem created by the fact that the other lenders have information that you do not have and how you tried to overcome this problem.

   b. Based on the market outcome and your realized profits or losses, your diagnosis of why you made or lost money, and how you modified your strategy in stage 2 based on this diagnosis.

   Be as precise as possible, for instance by including the actual mathematical formulas you use, if your methodology allows it.

Your input for the second stage is due on **November 3rd** at 24:00 by email to hombert@hec.fr.

Evaluation

The evaluation of this work will be based on two criteria:

- Performance in the game, i.e., total profits made evaluated both in absolute term (the level of profits) and in relative term (ranking relative to other groups): 10% on the first stage, 20% on the second stage.

- Quality of the report: 70%.

- Bonus points will be attributed to the team with the funniest name.

Rules

For this game to make sense and have pedagogical value, it is important that you do not see or use data from the other groups, and that you do not talk with other groups about how to set your prices. I will check statistically for "odd" forecasts or pricing behaviors. You can talk with the other groups about other aspects of the game (making sure you understand the rules, which statistical methods to use, etc.). You can also ask us questions on Slack.

Feedback from last year

Last year's students played a similar game for the group assignment of this course. We provided feedback on the game to the students. To help you with the assignment, we are also giving you this feedback from last year (see below).

Good luck with the game and have fun ☺

Feedback from last year

Thank you all for playing our lending game! All groups did a good job, some an outstanding one. This type of game can be completely spoiled if even only one participant behaves in a "crazy" way (say, gives zero interest rates to all borrowers), so it's important that everyone makes their best effort at finding a good strategy. This is what happened, and as a result there's plenty to learn from the outcome of the game, actually both for the students and the instructors. We summarize below a couple of points we find important.

Statistical approach

You used many different methodologies: simply partitioning the borrowers according to some observed correlations, logistic regressions, nearest neighbor, random forests, etc.

A critical methodological choice was whether to forecast a default probability for each borrower, or to classify borrowers into likely defaulters vs. unlikely defaulters.

The groups following the second approach tend to perform less well, for several reasons. First, as some realized, it is statistically very difficult to classify borrowers into these two categories, because default is rare and noisy. Imagine for instance that the correct model is that some borrowers have a 20% probability of default and others a 10% probability of default. If you try to classify them into "likely defaulters" vs. "unlikely defaulters" the best prediction is to consider all of them as unlikely defaulters, which is not a very informative model.

Second, from an economic perspective it also made more sense to forecast a probability of default. Imagine a borrower with a probability of default of 40%. If you charge an interest rate high enough to compensate for the default risk (in this case, 67%, see below), it is actually profitable to lend to this borrower. So the goal was not really to avoid bad borrowers, but rather to charge the appropriate interest rate for each borrower. This was possible with different approaches, but those that were more "continuous" were probably better designed for setting the interest rate appropriately.

Break-even interest rate

It was a good idea to start by computing the "break-even interest rate" $\overline{r}_i$, i.e., the minimum rate at which you should accept making a loan to a given loan applicant i. Assume you know the default probability PD_i of borrower i. Then if you lend to this borrower you will disburse 10,000 euros immediately, and will get $10,000 \times (1 + \overline{r}_i)$ later with probability $1 - PD_i$. Hence, your expected profit when lending to this borrower is:

$$-10,000 + 10,000 \times (1 + \overline{r}_i) \times (1 - PD_i)$$

The break-even rate is such that this expected profit is zero, which gives:

$$\overline{r}_i = \frac{PD_i}{1 - PD_i}$$

Several groups arrived at this formula through various means. Conversely, some groups arrived at poor decisions because they computed their expected profit wrong. In some cases this generated extremely low interest rates, leading to a large market share and significant losses.