Advanced Finance

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 scrapped from social media platforms. Because different lenders use different scrapping algorithms, they obtain different private signals capturing different dimensions of loan applicants' activity on social media. 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:

- id: unique individual identifier
- sex: 1 = male, 0 = female
- marital: 1 = married, 0 = other
- employment: employment status (four categories)
- income: annual income in euro (top coded at 1M euros)
- default: 1 = the borrower defaulted on the loan, 0 = the loan was repaid

- social1: social media activity coded from 0 to 1 as measured by lender 1
- social2: social media activity coded from 0 to 1 as measured by lender 2
- social3: social media activity coded from 0 to 1 as measured by lender 3

Note that past loans data contain the social media 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, marital, employment and income in their database, each lender only observes its own social media 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 four types of loan applicants:

- Type 0 (70% of the loan applicants): They always choose the cheapest offer. Formally, they choose the lowest among $r(1, i)$; $r(2, i)$; and $r(3, i)$.
- Type 1 (10% 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 (10% 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 (10% 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 0 applicant takes lender 1's offer at 5%.
- 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, 10% of the 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. Your cost of capital (discount rate) is 0% and your operating costs are zero too. 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

Team formation. The game requires not only finance skills but also statistical and programming skills. We therefore encourage you to form teams of students with different backgrounds (including different nationalities) and diverse sets of skills. Teams may be composed of three or four students. Please send the composition of your team by one week after the first class to hombert@hec.fr and cc all the team members. We will send you the data after you communicate us the composition of your team. If you have difficulties finding teammates, please email us and we'll help you with this.

The game 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. To predict default, you need to estimate a model of default using the past loans data. A simple way to do this, which you should try to implement in your group work, is to estimate an ordinary least square (OLS) regression explaining the event of default.[1] Standard statistical softwares and programming languages have built-in functions implementing OLS (or other statistical models).

NB: When you want to use OLS, you first have to choose the regression specification you want to run. For example, one very simple specification would be:

$$\text{default} = \alpha + \beta_1 \,\text{sex} + \beta_2 \,\text{marital} + \beta_3 \,\text{income} + \beta_4 \,\text{social1} + \epsilon$$

Running this model with OLS would give you estimates for the coefficients α, β_1-β_4. For example, the β_1 coefficient tells how much more (or less) likely is default for a male lender as compared to a female one. Of course, you want to experiment a bit with choosing your model. For example, you may want to try out using as explanatory variables the employment categories, the other social media scores, non-linear versions of some of the variables (e.g., squared or log income), or interactions between some of the variables (e.g., log income, marital status, and the product between the two). When choosing between different models, you should look at which explanatory variables have the greatest predictive power as indicated by their t-statistic (the higher the t-statistic on a given explanatory variable the higher its predictive power).

Once you have chosen and estimated your scoring model, you must decide on the interest rate you offer to each new loan applicant. Your goal is to maximize your total profit.

In stage 1, you are asked to:

- Choose a name for your fintech. Indicate in the email the composition of your team and the name of your fintech.
- Submit a csv file with the list of the 100,000 loan applications (identified by the variable "id") and the interest rate you offer to each applicant (call this variable "rate"). 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 empty for this applicant.

Your input for the first stage is due on **April 24** at 24:00 by email at hombert@hec.fr. You can start working immediately on the data and the scoring model. However, you should wait until

[1] If you master more advanced techniques, you can estimate a logistic model or even machine learning (ML) methods such as random forests. ML can accommodates non-linear effects of explanatory variables as well as interactive effects between several explanatory variables. It also parsimoniously selects the explanatory variables with the greatest predictive power.

the third class to start working on setting the interest rate because we will study the theoretical underpinnings required to make this business decision in the third class.

Stage 2. At the end of stage 1, we 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. We 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:

- Submit a csv file with the list of the 100,000 new loan applications, again with the interest rate you offer to each applicant.
- 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 **May 8** at 24:00 by email at 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%.

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. We 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.).