

Digital Technologies and Value Creation (Philippe Blaettchen) – Individual Assignment for Week 7

The Business Context

A major bank wants to better predict the likelihood of default for its customers, as well as identify the key drivers that determine this likelihood. They hope that this would inform the bank's decisions on who to give a credit to and what credit limit to provide, as well as also help the bank have a better understanding of their current and potential customers, which would inform their future strategy, including their planning of offering targeted credit products to their customers.

The Data

The bank collected data on 25 000 of their existing clients. Of those, 1,000 were randomly selected to participate in a pilot described below. Data about the remaining 24,000 is in the file "Credit_data.csv". The dataset contains various information, including demographic factors, credit data, history of payment, and bill statements from April to September, as well as information on the outcome: did the customer default or not in October.

Data Dictionary

- **ID:** ID of each client
- **LIMIT_BAL:** Total amount of credit line with the bank (including all individual and family/supplementary credit)
- **SEX:** Gender (1=male, 2=female)
- **EDUCATION:** Education (1=graduate, 2=undergraduate, 3=high-school, 4=other, 0,5,6=unknown)
- **MARRIAGE:** Marital status (1=married, 2=single, 3=other, 0=unknown)
- **AGE:** Age in years
- **PAY_1:** Repayment status 1 month ago – in September (-2=no need to pay, zero balance, "payment holiday", etc., -1=paid in full, 0=revolving credit, i.e., client paid more than the minimum payment, but less than the total balance, 1= delay for one month, ... 8=delay for 8 months, 9=delay for 9 months or more)
- **PAY_2:** Repayment status 2 months ago – in August (scale as above for PAY_1)
- **PAY_3:** Repayment status 3 months ago (scale as above for PAY_1)
- **PAY_4:** Repayment status 4 months ago (scale as above for PAY_1)
- **PAY_5:** Repayment status 5 months ago (scale as above for PAY_1)
- **PAY_6:** Repayment status 6 months ago (scale as above for PAY_1)
- **BILL_AMT1:** Amount of bill statement 1 month ago – in September
- **BILL_AMT2:** Amount of bill statement 2 months ago
- **BILL_AMT3:** Amount of bill statement 3 months ago
- **BILL_AMT4:** Amount of bill statement 4 months ago
- **BILL_AMT5:** Amount of bill statement 5 months ago
- **BILL_AMT6:** Amount of bill statement 6 months ago
- **PAY_AMT1:** Amount of payment 1 month ago – in September
- **PAY_AMT2:** Amount of payment 2 months ago
- **PAY_AMT3:** Amount of payment 3 months ago

- **PAY_AMT4:** Amount of payment 4 months ago
- **PAY_AMT5:** Amount of payment 5 months ago
- **PAY_AMT6:** Amount of payment 6 months ago
- **Default_0:** Default in October (1=yes, 0=no)

Your Pilot Project

Your department wants to pilot a new product, a short-term credit line with the limit of £25,000, and for the purposes of this assignment assume that the line is for 1 month at 2% per month. More so, assume that the client who was issued credit and repaid it will more likely use your bank for similar short-term financing needs in the future, which has an additional lifetime value (CLV) of 1,000. However, if the client will default, then you will be able to recover only 20,000 out of 25,000 granted.

The data about the 1,000 clients that were randomly selected for this pilot is in the file "New_applications.csv".

The ultimate question: which of the 1,000 "new applicants" in the pilot should be issued credit?

In your analyses, please make the following simplifying assumptions:

1. Defaults on the previously issued credit is not your problem
2. All the clients who will be offered the credit line will use it in full
3. Your cost of capital is 0

In other words, for each client in the pilot, if the credit is issued and repaid, then the bank earns a profit of $25,000 \times 2\% + 1,000 = 1,500$; if the credit is granted but the client defaults, then the bank loses $25,000 - 20,000 = 5,000$; and if the credit is not issued, then profit = loss = 0.

The Assignment

All questions are to be answered in the corresponding section within the quiz. In addition, please make sure to upload your code.

1. Determine which of the 1,000 clients in the pilot should be issued credit. Once done, create a spreadsheet with only one column, A1:A1000, of 0s and 1s, representing your recommendation for issuing credit to each of 1 000 pilot customers in the order of their IDs as per the data (1 – issue credit, 0 – do not issue).

Copy-paste this column into the answer field below. [Please paste 1000 digits, 1 or 0, with no other symbols (no commas, dashes, or anything else). Each digit should appear either on its own line, or there should be a space between each digit. Please make sure that your answer appears in either of those formats.

Example a: 1 0 0 1 ...

Example b:

1
0
0

1
...]

How did you set your threshold? Did you engineer any extra “features”? We will calculate the profit that the bank would have actually received following your recommendation (We can calculate this because we have the data about which of these 1,000 clients will actually default and which will not – you do not have this data). Your score for this part of the assignment will be based solely the profits you achieve.

2. Which three of the 1,000 pilot clients are most likely to repay the loan if it were granted to them? [Please paste 3 IDs with spaces between them (no commas, or anything else). Example: n1000-1 n1000-88 n1000-69]
3. Which three of 1,000 pilot clients are least likely to repay the loan if it were granted to them? [Please paste 3 IDs with spaces between them (no commas, or anything else). Example: n1000-1 n1000-88 n1000-69]
4. A colleague came by to discuss two clients: A and B. The colleague says that a model predicts that A will repay and B will default. What should the bank do?
 - a) Issue credit to A and do not issue credit to B
 - b) Issue credit to A, but do some more work on B
 - c) Do not issue to B, but do some more work on A
 - d) The statement above does not provide enough information to determine the best course of action.

If you choose responses b-d, you will need to describe what exactly you suggest needs to be done to make a decision.

As we discussed in class, models predict probabilities, which are then compared to some threshold, T , and the prediction is classified as positive (“yes”) if $\text{Prob} > T$, and it is classified as negative (“no”) otherwise.

6. As T increases, what will happen to the confusion matrix and its metrics:
 - a) To total number of correct “yes” predictions will increase
 - b) Sensitivity (i.e., the percentage of correct “yes” predictions) will increase
 - c) The total number of “yes” predictions will increase
 - d) Specificity (i.e., the percentage of correct “no” predictions) will increase
7. As T increases, what will happen to the ROC curve:
 - a) It will be shifting toward the lower-left
 - b) It will be shifting toward the upper-right
 - c) It will not be shifting
 - d) It will be shifting, but it is impossible to predict in which direction (depends on the data and/or the model)

8. Consider the situation where we are predicting yes/no labels for our data at random (i.e., there is a 50% chance of picking “yes” and a 50% chance of picking “no”). Assume that we pick a threshold strictly between 0 and 1. Explain why, whatever this threshold, the point we obtain on the ROC curve is $(1/2, 1/2)$.

A few hints:

- Do not over-think – start with an MVP (“minimum viable product”). Use the data “as is,” apply minimal data cleaning and pre-processing, build a simple predictive model, and use its results to go “all the way” to arrive to the final issue/no issue decision for each new pilot applicant. If you’re using logistic regression, you will likely have to increase the number of iterations – e.g., `LogisticRegression(max_iter=...)` instead of `LogisticRegression()`.
- Think about how you can use the existing 24,000 clients data to determine (and evaluate) a strategy for deciding which of the new 1,000 should be issued credit once you have the model predictions for them.
- Once you have a working MVP model on the “as is” data, consider feature engineering. Think about what information is contained in your data but is not currently captured by your variables. Brainstorm new variables that you can “engineer” from your data, create them and add to your data, rerun your models and comment on the improvements in the model’s predictions. Creative feature engineering will improve your model accuracy and will allow you to obtain a higher profit.