**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 “MMA867 A3 – credit data.xls”. 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.

The screenshot below depicts the first 10 rows of the data:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | LIMIT\_BAL | SEX | EDUCATIO | MARRIAGE | AGE | PAY\_1 | PAY\_2 | PAY\_3 | PAY\_4 | PAY\_5 | PAY\_6 | BILL\_AMT1 | BILL\_AMT2 | BILL\_AMT3 | BILL\_AMT4 | BILL\_AMT5 | BILL\_AMT6 | PAY\_AMT1 | PAY\_AMT2 | PAY\_AMT3 | PAY\_AMT4 | PAY\_AMT5 | PAY\_AMT6 | default\_0 |
| 1 | 20000 | 2 | 2 | 1 | 24 | 2 | 2 | -1 | -1 | -2 | -2 | 3913 | 3102 | 689 | 0 | 0 | 0 | 0 | 689 | 0 | 0 | 0 | 0 | 1 |
| 2 | 90000 | 2 | 2 | 2 | 34 | 0 | 0 | 0 | 0 | 0 | 0 | 29239 | 14027 | 13559 | 14331 | 14948 | 15549 | 1518 | 1500 | 1000 | 1000 | 1000 | 5000 | 0 |
| 3 | 50000 | 2 | 2 | 1 | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 46990 | 48233 | 49291 | 28314 | 28959 | 29547 | 2000 | 2019 | 1200 | 1100 | 1069 | 1000 | 0 |
| 4 | 50000 | 1 | 2 | 1 | 57 | -1 | 0 | -1 | 0 | 0 | 0 | 8617 | 5670 | 35835 | 20940 | 19146 | 19131 | 2000 | 36681 | 10000 | 9000 | 689 | 679 | 0 |
| 5 | 50000 | 1 | 1 | 2 | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 64400 | 57069 | 57608 | 19394 | 19619 | 20024 | 2500 | 1815 | 657 | 1000 | 1000 | 800 | 0 |
| 6 | 1.00E+05 | 2 | 2 | 2 | 23 | 0 | -1 | -1 | 0 | 0 | -1 | 11876 | 380 | 601 | 221 | -159 | 567 | 380 | 601 | 0 | 581 | 1687 | 1542 | 0 |
| 7 | 140000 | 2 | 3 | 1 | 28 | 0 | 0 | 2 | 0 | 0 | 0 | 11285 | 14096 | 12108 | 12211 | 11793 | 3719 | 3329 | 0 | 432 | 1000 | 1000 | 1000 | 0 |
| 8 | 20000 | 1 | 3 | 2 | 35 | -2 | -2 | -2 | -2 | -1 | -1 | 0 | 0 | 0 | 0 | 13007 | 13912 | 0 | 0 | 0 | 13007 | 1122 | 0 | 0 |
| 9 | 2.00E+05 | 2 | 3 | 2 | 34 | 0 | 0 | 2 | 0 | 0 | -1 | 11073 | 9787 | 5535 | 2513 | 1828 | 3731 | 2306 | 12 | 50 | 300 | 3738 | 66 | 0 |
| 10 | 260000 | 2 | 1 | 2 | 51 | -1 | -1 | -1 | -1 | -1 | 2 | 12261 | 21670 | 9966 | 8517 | 22287 | 13668 | 21818 | 9966 | 8583 | 22301 | 0 | 3640 | 0 |

# 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, 5,6=unknown)
* **MARRIAGE**: Marital status (1=married, 2=single, 3=other)
* **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 (meaning 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)

# 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 credit granted.

The data about 1 000 clients that were randomly selected for this pilot is in the file “MMA867 A3 - new applications.xlsx». The screenshot below depicts the first 5 rows of the data:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | LIMIT\_BAL | SEX | EDUCATIO | MARRIAGE | AGE | PAY\_1 | PAY\_2 | PAY\_3 | PAY\_4 | PAY\_5 | PAY\_6 | BILL\_AMT1 | BILL\_AMT2 | BILL\_AMT3 | BILL\_AMT4 | BILL\_AMT5 | BILL\_AMT6 | PAY\_AMT1 | PAY\_AMT2 | PAY\_AMT3 | PAY\_AMT4 | PAY\_AMT5 | PAY\_AMT6 |
| n1000-1 | 5.00E+05 | 1 | 1 | 2 | 29 | 0 | 0 | 0 | 0 | 0 | 0 | 367965 | 412023 | 445007 | 542653 | 483003 | 473944 | 55000 | 40000 | 38000 | 20239 | 13750 | 13770 |
| n1000-2 | 210000 | 1 | 1 | 2 | 29 | -2 | -2 | -2 | -2 | -2 | -2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| n1000-3 | 150000 | 1 | 1 | 2 | 27 | 0 | 0 | 0 | 0 | 0 | 0 | 86009 | 86108 | 89006 | 89775 | 87725 | 40788 | 4031 | 10006 | 3266 | 4040 | 1698 | 800 |
| n1000-4 | 20000 | 1 | 2 | 1 | 38 | 0 | 0 | 0 | 0 | 0 | -1 | 17973 | 19367 | 19559 | 18240 | 17928 | 150 | 1699 | 1460 | 626 | 1750 | 150 | 0 |
| n1000-5 | 4.00E+05 | 1 | 2 | 1 | 34 | -1 | -1 | -1 | -1 | -1 | -1 | 19660 | 9666 | 11867 | 7839 | 14837 | 7959 | 9677 | 11867 | 7839 | 14837 | 7959 | 5712 |

# 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:

* Defaults on the previously issued credit is not your problem
* All the clients who will be offered the credit line will use it in full
* Your cost of capital = 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\*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 the profit=loss=0.

# Questions (1,2,3):

# 1: Issuing credit [60pts]. Determine which of the 1000 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, 0 do not issue). Name this spreadsheet with your team name (e.g., “Team Ay.xls”) and upload it on portal.

# 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:

* + Profit <=0 0 points
  + Profit >0, but <=100,000 20 points
  + Profit >100,000, but <=300,000 30 points
  + Profit >300,000, but <=500,000 40 points
  + Profit >500,000, but <=600,000 50 points
  + Profit >600,000 60 points

**2: Writing the report [10pts].** In addition to the spreadsheet that simply contains the recommendations please also prepare a PDF report describing how you approached the task.

As you are working on the report, consider the following sub-questions:

* Which three of 1 000 pilot clients are most likely to repay the loan if it were granted to them?
* Which three of 1 000 pilot clients are least likely to repay the loan if it were granted to them?
* 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?
  + Issue credit to A and do not issue credit to B
  + Issue credit to A, but do some more work on B
  + Do not issue to B, but do some more work on A
  + The statement above does not provide enough information to determine the best course of action.
* 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 Prob>T, and it is classified as negative (“no”) otherwise.
* As T increases, what will happen to the confusion matrix and its metrics:
  + The total number of correct “yes” predictions will increase
  + Sensitivity (i.e., the percentage of correct “yes” predictions) will increase
  + The total number of “yes” predictions will increase
  + Specificity (i.e., the percentage of correct “no” predictions) will increase
* As T increases, what will happen to the ROC curve
  + It will be shifting toward the lower-left
  + It will be shifting toward the upper-right
  + It will not be shifting
  + It will be shifting, but it is impossible to predict in which direction (depends on the data and/or the model)

There is no need to answer these questions in the report but be ready to discuss them in class.

**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

Think about how you can use the existing 24 000 client’s 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 allow you obtain a higher profit for Q1.

**3: Exploring the notion of “Responsible AI” [30pts].** Equality and anti-discrimination are among the major concerns of the modern-day society. There are numerous social movements, regulations, business practices, and, generally speaking, lots of “attention” devoted to these issues. Up until very recently, however, this attention was almost solely focused on human decision makers, e.g., how do we ensure that (some) humans do not discriminate against (other) humans. Discrimination is, of course, a subset of the broader notion of “ethics” – a set of principles that govern “good” behaviors and decision-making.

The arrival of artificial intelligence adds a new twist to discrimination and ethics. Indeed, as more and more decisions are made by machines, we must ask ourselves how we should change our thinking about the ethical principles guiding those decisions, anti-discrimination being one of them?

In the last question of the assignment, you will explore this fundamentally important question by looking at the role of gender in credit provision by a financial institution.

**[Very Brief] Overview of Anti-discrimination Legislation in Financial Services**

Many leading economies instituted a variety of policies to fight discrimination. In the USA, for example, the Equal Credit Opportunity Act (ECOA, 1974) prohibits financial institutions from collecting data about the characteristics such as race, colour, national origin, gender, marital status, religion, receipt of public assistance, or exercise of consumer protection rights; clearly such characteristics (because they are not collected) cannot be used in credit scoring either.

In the EU, these practices are governed by Articles 8 and 19 of the Treaty of the Functioning of European Union (TFEU), as well as the so-called “Gender Directive”, Council Directive 2004/113/EC of December, 13 2004, on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation. With respect to using gender in determining the access to and supply of goods and services, these regulations historically permitted the use of gender data, so long as it could be justified by the underlying actuarial and statistical data. However, in March 2011 the European Court of Justice banned the use of gender in various financial and risk decision-decision making models. The reaction to this ruling has been mixed: while some called it a “key victory for women’s rights, say civil society organisations”, others noted that “the decision … to ban gender in [our product] pricing was disappointing as it could have an impact on our customers”. The EU regulations, do not, however, prohibit the collection of the gender data (just its use in training models)

In Canada, the use of data in algorithmic decision-making is not as clearly regulated as in some other countries. Anti-discrimination is covered in the Canadian Human Rights Act, but practically speaking, there is a lot of “wiggle room.” In this situation, the industry is in a unique position to lead and shape the Ethics and AI discussion, and yours truly Smith School of Business is at the forefront of this activity, e.g., see <https://smith.queensu.ca/centres/scotiabank/events/event/ethics-ai.php>

To explore the roles/impacts of these regulations, do the following:

**[5 pts]:** Rerun your best model from Q1 with and without the use of the “SEX” variable. Comment on the resultant predictive performance of the two models.

**[10pts]:** Now apply these two models (i.e., with and without gender) separately to males and females. What do you observe?

**Hint**: it may be insightful to approach this question as follows. Ultimately, and as you saw in Q1, while the predicted object is a probability, the company needs to make a binary decision: give the credit or not. To do so they will set a cut-off (threshold) value (e.g., 20%) and if the probability is below that, then give, and otherwise not. Thus, create a graph that, as a function of this threshold, depicts the percentage of males and females getting credit if the company used one model versus the other.

**[15 pts]:** What are the implications of your analyses for the debate on equality/anti-discrimination/ethics and algorithmic decision-making/AI? Be specific: if you suggest something, support it with the analyses.

**Hint**: What can a firm do under the EU regulation but not under the US regulation, and how does that impact your results above (in particular, the impact of removing the SEX variable on credit provision between genders depicted on the graph you created in the previous question)