STAT 628 Credit Risk Project: Installment 2

fall 2025

Problem Introduction

You have been asked to help a local credit union assess the credit risk of potential borrowers who have applied for small business loans. The credit union tends to attract borrowers who have started small local business like cafes, dry cleaners, or bodegas but lack extensive credit histories or background. Typically, these borrowers, having exhausted most of their personal funds to start their businesses, seek short-term loans to help keep their businesses running. However, their lack of credit background makes these entrepreneurs too risky for large national banks. And so these potential borrowers have turned to the local credit union and applied for a loan.

The credit union issues loans with 12 month terms. That is, it expects to recover the principal and interest within one year. The principals range from about \$10,000 to \$2 million. To obtain payments from the borrower, the credit union garnishes a fixed percentage of the credit card transactions processed by the borrower. The credit union extracts this percentage until it recovers not only the principal but also the interest. That is, each month the lender take a cut from all credit card transactions made by the business. The percentage is fixed at the time that the loan originates and is set in anticipation that the borrower can pay off the loan in 12 months.

Of course, there is monthly variation in credit card transactions: in some months, businesses might have a large number of sales and in some months, they might not have many sales. Such fluctuations contribute to the risk the credit union assumes when it lends money. If a business enjoys strong sales for several months, the credit union may recover its money (plus interest) quickly. On the other hand, if a business experiences a prolonged period of reduced sales, the credit union is unlikely to recover its money within the year.

An example

As an example, suppose that the owner of a new cafe obtains a \$20,000 loan from the credit union. With interest of 20%, the cafe owner must repay a total of \$24,000 to the credit union within 12 months. To ensure on-time repayment, the credit union wants to recover \$2,000 per month by garnishing the cafe's credit card transactions. When the cafe owner applied for the loan, they reported that, on average, they processed \$15,000 per month in credit card transactions. Based on this information, the credit union decided to garnish $2000/15000 \times 100\% \approx 13.3\%$ of all credit card transactions processed by the cafe.

If the cafe has a strong month of sales and processes \$22,500 in credit card transactions, the credit union will receive \$3,000, putting the cafe ahead of its payment schedule. On the other hand, if the cafe experiences a slow month and only processes \$12,000 in transactions, the credit union will receiving only \$1,600, putting the cafe behind its payment schedule.

Your task and description of the data

The credit union has created a new metric of credit worthiness called **P**erformance **R**atio at **S**ix **M**onths (`PRSM`). The PRSM score is computed as

$$\text{PRSM} = 2 \times \frac{\text{Amount repaid at 6 months}}{\text{Total amount owed}}.$$

In an ideal world, every borrower would have a PRSM score of 1, indicating that they have paid back exactly half the total amount owed at six-months. Of course, the world is far from ideal and variation in each borrower's monthly transactions lead to variations in PRSM scores. Generally speaking, PRSM scores greater than 1 indicate that the borrow is "ahead of schedule", in the sense that they have paid off more than half the total amount at six months. As you saw in installment 1, the credit union has historically had many borrowers with PRSM less than 1. The credit union would like to do a better job in identifying these potentially risk loans. To do so, it has collected lots of data about loans they have given in the past. You have received a dataset derived from a random sample of these past loans containing the following variables:

- The PRSM score of the loan (`PRSM`).
- The Fair Isaac Credit Score (FICO) of the borrower (`FICO`). FICO scores generally range from 300 to 850. Most credit agencies classify FICO scores into five categories: Poor ($300 \leq$ FICO ≤ 579), Fair ($580 \leq$ FICO ≤ 669), Good ($670 \leq$ FICO ≤ 739), Very Good ($740 \leq$ FICO ≤ 799), and Excellent ($800 \leq$ FICO ≤ 850). See here for more information about FICO score.
- The total amount owed to the credit union (`TotalAmtOwed`). This amount reflects the principal of the loan as well as all interest. The credit union wants to recover this amount from the borrower at the end of the 12 months.
- Expected volume of credit card transactions per month (`Volume`, in $)
- Ratio of the monthly garnishment to the expected volume of credit card transactions (`Stress`).
- Number of delinquent credit lines (`Num_Delinquent`). Delinquency occurs when a business is more than 30 days behind payment of a debt.
- Total number of credit lines (`Num_CreditLines`), including both delinquent and non-delinquent lines.
- An indicator of whether the business is owned by a woman (`WomanOwned` is 1 if woman-owned and 0 otherwise).
- A categorical predictor (`CorpStructure`), which records whether business is structured as a sole proprietorship, corporation, limited liability corporation (LLC), or a partnership.
- 6-digit NAICS code (`NAICS`). The North American Industry Classification System provides a 5- or 6-digit code that classifies different industries. For instance, the code for universities and colleges is 611310. You can look-up individual codes at this link.
- The number of months for which the business has been open (`Months`).

You will build a multiple regression model to predict `PRSM` using the provided predictors (and possibly additional predictors derived from those provided). You will receive two datasets, a **training** dataset, which contains measurements of `PRSM` and all the predictors, and a **evaluation** dataset, which contains measurements of all predictors but not `PRSM`. You will use the training dataset to fit your model (i.e., estimate model parameters, perform inference, and check relevant model diagnostics). You will then use the fitted model to predict `PRSM` for each loan in the evaluation dataset (using, e.g., the `predict()` function).

Deliverables

You will upload 5 documents to Canvas: presentation slides (due first); a 2-page `pdf` document with your executive summary, references & contributions; a technical report that includes all relevant code (one `pdf` file and one `rmd` or `qmd` source file); and point and interval predictions for the loans in the evaluation dataset (one `csv` file). In addition to these items, you will give an in-class presentation.

Executive summary

> You should submit a 2-page document with your executive summary on page 1, and your reference & contributions on page 2.

The executive summary should present your conclusions and be free of technical jargon, figures, graphs, and code. In other words, you should describe and interpret your *results* and should *not* describe the process by which you obtained your results. The executive summary should convey the main conclusions of your modeling efforts in language that is accessible to someone who may not have taken STAT 628 before. Your executive summary should

- List the factors that drive substantial variation in PRSM. You should **not** exhaustively list all variables that have a statistically discernible effect. It is possible that a predictor has a statistically discernible effect but not a practically relevant (significant) one.
- Introduce a baseline potential borrower by selecting values of each predictor in your final model. Report the predicted PRSM (along with relevant uncertainty intervals) of your baseline borrower.
- Describe the main drivers of PRSM and indicate which are associated with greater or lesser credit risk relative to the baseline.
- Use conveniently rounded numbers when forming a baseline borrower and highlighting changes relative to that baseline. Similarly, 1 unit changes in some predictors may not be that relevant; consider using more realistic or practically relevant changes.

Your executive summary should not be longer than 1 page. Do not adjust the page margins or use an extremely small font size to achieve this. It should be free of technical jargon and make minimal use of equations.

Your executive summary will be graded on a 0–5 point scale based on the extent to which it delivers the above requested items in a concise and clear fashion.

On page 2, you should list any references that you used, and list your group member contributions to each aspect of the project, as illustrated in the example table below (but feel free to organize contribution categories as you feel is most appropriate).

Contributions	Cecile Ane	John Doe	Jane Doe
code & analysis	responsible for data cleaning. reviewed data analysis & interpretation.	responsible for figures, model choice. reviewed and provided feedback on all steps.	responsible for analysis & results. reviewed and gave feedback on all steps
summary & report write-up	responsible for introduction, data cleaning explanations, conclusion & refs.	responsible for good figures. reviewed/edited both documents. checked reproducibility.	responsible for first draft of summary. reviewed/edited report.
presentation	responsible for slides 1-4, reviewed/edited slides 5-10	responsible for slides 5-10, reviewed/edited slides 1-4	reviewed/edited all slides

Unless otherwise noted, all group members will receive the same grade for the project. However, if we find that there were serious, unequitable contributions or there was a lack of regular communication between team members, we will assign points individually within a group. Please remember that this may also mean that some group members who may have contributed the most to the project may still lose points if they didn't make an earnest attempt to keep others informed about their progress or they weren't being a "team player" with respect to sharing the workload.

Technical report and reproducible code

The technical report is meant to convey to your peers that your analysis was sound. It is not, however, meant to be a step-by-step chronology of what you did to arrive at your final model. Instead, you should summarize the most important steps of your modelling process. At the very least, you should describe and justify (when necessary) the following:

- the removal of any outliers or suspect observations,
- all transformations and the construction of any new predictors, and
- the procedure used to select the final model. If you use an automated procedure like stepwise regression or the LASSO, your text needs to provide enough detail so that someone reading your technical report could reproduce (re-code) your findings (without access to your code).

Your technical report must include the following:

- the printed R summary of your fitted model,
- diagnostics to assess the extent to which the usual multiple regression model assumptions hold (include any relevant graphics),
- term-by-term and estimate-by-estimate interpretations of the parameters included in your final model,
- an explanation for how you selected the characteristics of the baseline borrower in your executive summary.

You are strongly encouraged to prepare your technical report using quarto or R markdown. While this allows you to include all relevant code, you should take care not to include excessive output in your report. You should also ensure that code and output do not extend beyond the page margin and that your figures are appropriately sized. For each plot that you include, you must explain its relevance in the exposition. See Section 5.3 of the R Markdown Cookbook for information about controlling page margins and Section 5.4 of the same book for information about sizing figures. The instructor and TA may run the code that you submit and will deduct marks if they are unable to reproduce the results of your analysis.

Your technical summary will be graded on a 0–5 point scale based on its clarity and style; the degree, the technical soundness of your analysis; and the extent to which it includes the requested items. The instructor and TA may also run the code provided in the technical summary. If they are unable to run all of the code successfully and reproduce all reported numerical summaries and visualizations, up to 4 (out of the 5) points will be taken off.

Predictions

Once you have fit your final model, you should load the evaluation data. If you created new predictors from the original ones, you will need to manually update the evaluation data to include these new predictors (in `R`, a function like `dplyr::mutate` is really helpful for this purpose) before you run your prediction, e.g. using the `predict` function. Save your predictions into its own local variable and then write that variable to a `csv` file in your working directory. The code chunk below contains a template for the relevant `R` code. Note that you will need to **adjust this code** as appropriate (to your model name, etc).

```
evaluation_data <- read_csv(file = "evaluation_data.csv")
# here: do any necessary transformation or create any additional needed predictors
predictions <- predict(
  object = fitted_model,
  newdata = evaluation_data,
  interval = "prediction"
)
write_csv(predictions, filename = "predictions.csv")
```

You will be asked to upload the `csv` file containing your predictions to Canvas alongside your technical report and executive summary. That `csv` file should list the loans in the same order in which they are listed in the input evaluation file `evaluation_data.csv`, and should have 3 named columns: one for the point prediction, one for the lower bound of the 95% prediction interval, and one for the upper bound of the 95% prediction interval.

You will receive a score between 0 and 5 for your predictions. By submitting the predictions, you will earn 2 points. The instructor and TA will compute the root mean square error (RMSE) of your predictions. Students with RMSE below the class median will earn 1 extra point (while students with RMSE larger than the class median will not). The instructor and TA will also compute the average coverage of your 95% prediction intervals. If your coverage lies between 80% and 90%, you will receive 1 point and if the coverage exceeds 90%, you will receive 2 points.

Presentation

You group will give a short, 10-minute presentation that (i) summarizes your final model and main conclusions about what drives credit risk; (ii) briefly describes your overall modeling approach (i.e., what you did and why); and (iii) outlines avenues for future development or improvement. Each group member is expected to speak during the presentation, but the overall division of speaking time need not be uniform.

Your presentation will be graded on a 0–5 point scale based on the content and organization of your presentation slides, the clarity and structure of your oral presentation, and your ability to answer audience questions.

Modeling hints

Discussions with the credit union has suggested potential issues, which could be helpful in modeling PRSM.

1. There is a strong belief that businesses that have been open longer are more credit-worthy. However, experience suggests that after a certain point, an additional month of operation has a diminished predictive effect.
2. There is a belief at the credit union that woman-owned businesses tend to be more stable and have more consistent monthly credit card transaction, making it more likely they pay off their loans on time.
3. The credit union has observed that corporations may be slower than other businesses at paying back their loans.
4. Lenders very often ask independent credit bureaus to conduct a credit check and assess the ability of a potential borrower to repay a loan. As part of the check, these bureaus may look at the raw FICO score or its category. The credit union believes that information contained in the FICO score may be relevant when dealing with certain borrowers but not others.
5. There is a belief that certain predictors affect PRSM in a non-linear fashion. Consider transforming some predictors or creating new predictors by squaring or cubing individual numerical predictors or taking ratios of existing ones.
6. You may also consider creating new predictors by "discretizing" numerical predictors. That is, you can convert a numerical predictor into a categorical predictor by binning certain values into a category. A great example of this are the FICO score categories.
7. The credit union recently overhauled its data collection practices but there is concern that there may be errors in some of its historical data. Pay particular attention to values outside the range of certain variables. In an ideal world, we should first check with the credit union before excluding any observation, but here we do not have the opportunity to ask for confirmation whether an observation was recorded prior or after the overhaul of their data collection practice.