STAT333 Project Installment 2

17 April 2023

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 (i) have started small local business like cafes, dry cleaners, or bodegas but (ii) do not have extensive credit histories or background. Typically, in order to start their businesses, these borrowers have exhausted much of their personal funds and they are seeking 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.

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 $\frac{2000}{15000} \times 100\% \approx 13.3\%$ of all credit card transactions processed by the cafe.

Of course, there is monthly variation in the volume of the cafe's credit card transactions. In some months, they might be doing more business than usual (i.e. generating more than $15,000 in credit card transactions) while in others they might be doing less. Such fluctuations contribute to the risk the credit union assumes when it lends money: if the borrower's business does really well, the credit union might recover the money ahead of schedule. But if the borrower's business struggles for an extended period of time, the credit union may not recover its money (with interest) within the year.

Your task and description of 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 one 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 one. 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've 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 ≤ 670), 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 (`Volume`)
- 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 == 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. Note that this variable should be treated as a categorical predictor.
- The number of months for which the business has been open (`Months`).

Your will build a multiple regression model to predict `PRSM` using the provided predictors (and possibly additional predictors derived from those provided). Your group will received 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 model diagnostics). You will then use the fitted model to predict `PRSM` for each borrower in the evaluation dataset (using, e.g., the `predict()` function).

Deliverables

There are three project deliverables: a one-page executive summary; a technical report; and point and interval predictions for the observations in the evaluation dataset.

Executive Summary

The executive summary should present your conclusions and be free of technical jargon, figures, graphs, and R 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 333 before. Your executive summary should

- List the factors that drive substantial variation in PRSM. You do not need to exhaustively list all variables that have, e.g., a statistically significant effect. It is possible that a predictor has a statistically significant effect but not a practically relevant 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 one page. Do not adjust the page margins or use an extremely small font size to achieve this.

Technical report

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
- The procedure used to select the final model. If you use an automated procedure like stepwise regression or the LASSO, you need to provide enough detail so that someone reading your technical report could reproduce your findings

Your technical report must include the following:

- The printed R summary of your fitted model.
- Diagnostics to assess the extent to which the usual MLM 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 should prepare your technical report using RMarkdown. While this allows you to include all relevant R 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 RMarkdown Cookbook for information about controlling page margins and Section 5.4 of the same book for information about sizing figures. **The instructor and TA will run the code that you submit and will deduct marks if they are unable to reproduce the results of your analysis.**

Predictions

Once you have fit your final model, you should load the evaluation data into R. If you created new predictors from the original ones, you will need to manually update the evaluation data to include these new predictors (a function like `dplyr::mutate` is really helpful for this purpose!) before you run `predict`. Save the output from the `predict` function 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.

```
evaluation_data <- read_csv(file = "evaluation_data.csv")
### Do any necessary transformation or create any additional
### predictors you need here
predictions <- predict(object = fitted_model, new_data = 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 have 3 columns, one each of the point prediction, the lower bound of the 95% prediction interval, and the upper bound of the 95% prediction interval.

Honors Credit Assignment

The credit union wants an easy-to-use application in which they can plug in values of different predictors from a potential borrower and output projections about their PRSM score. Students taking the course for honors credit will create a Shiny web application that provides **live, user-friendly, interactive** predictions from the final, fitted model. At a minimum, your application should allow a user to input different X values (possibly as text-entry or via a drop-down menu) and output (i) the predicted PRSM score; (ii) a prediction interval; and (optionally) (iii) the probability that the PRSM score will exceed 1.

To receive honors credit, you must submit (i) the Shiny app code and (ii) any additional files that are required to run the Shiny app. The application must "compile" locally without any errors for grading. You must submit these deliverables (via email to the instructor) on the same day as the other deliverables are due.

Also, while you are highly encouraged to look at other examples of Shiny online, you may not use resources in a manner inconsistent with the academic integrity guidelines stated in the syllabus. If you relied on external examples of Shiny apps to construct yours, please submit a list of references, complete with links to the applications you consulted.

Modeling Hints

Discussions with the credit union has suggested several subtle 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 are much 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 should 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.