MXN600 Major Data Analysis Project

David Warne

19 September 2022

Introduction

This document describes your MXN600 Major Data Analysis Project. This assignment is to be completed as a group and it is expected that all group members contribute. This project consists of a substantial data analysis task using real (but simplified) credit risk data. Your group will:

1. Perform the analysis using GLMs and GLMMs and R,
2. Write a technical report with Rmarkdown,
3. Write a summary on a page (SOAP) document (can be 2 A4 pages), and
4. Give a presentation to the class on your findings to date in Week 12.

See the remaining sections for more details. Please contact me if you have any questions.

Scenario

You are a team of data analysts working for a peer-to-peer personal lending start up company that has recently been acquired by a regional Australian bank. Some senior financial analysts at the bank have some concerns about the credit risk models your company has been using. They have based their concerns on performance benchmarks (Figure 1) that they generated with your companies' model using some historical lending data available from the USA between 2007 to 2011. They concluded that these models are too ad-hoc to be suitable for use in a bank that is subject to strict regulatory requirements.

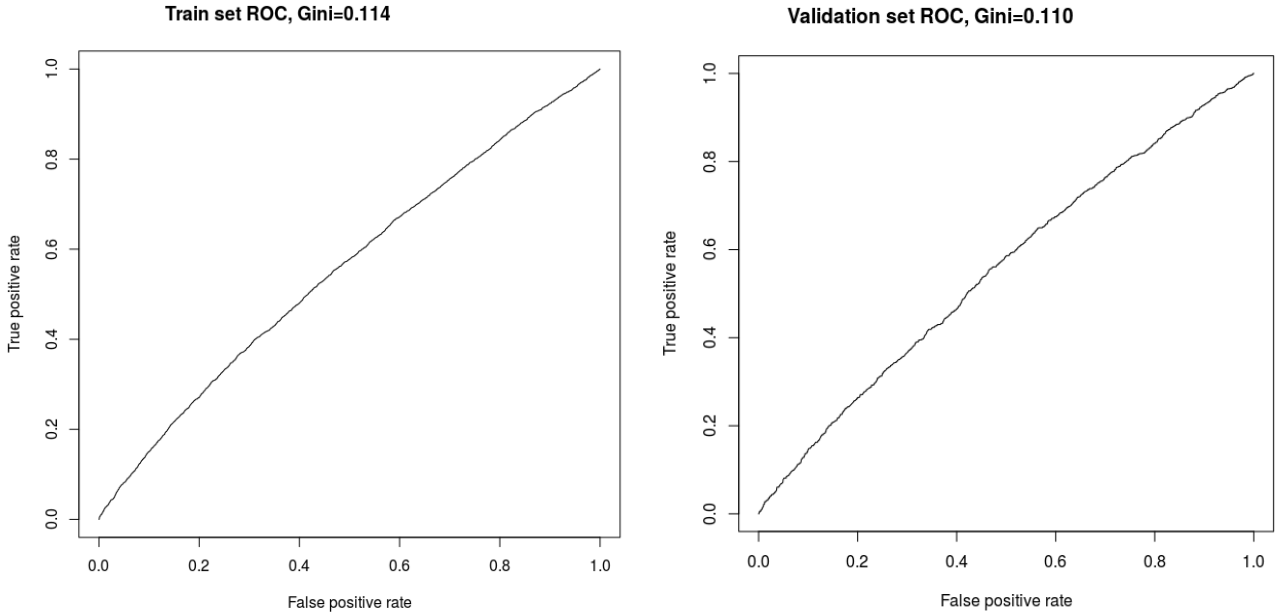

Figure 1: Receiver Operating Characteristic (ROC) curve and Gini score for your old model under a training set (Left) and a validation set (Right). The Gini score can be obtained from the area under the curve (AUC) using Gini = 2(AUC - 1/2).

Your new management at the bank has commissioned your team to do a 'ground-up' rebuild of your credit risk (loan default) model using tools, methods and performance measures that you are familiar with. The objective is to build a statistical model to **predict loan default based on information known at the time of**

loan application.

The bank's IT team has provided your team with the historical lending data that had been used to benchmark your old model (data for over 38,000 loans). This data is already split into a training and validation set. You will utilise these data to build your new credit risk model. There is also a data dictionary explaining each variable (20 possible covariates).

You have also been provided with an *extended version* of the data that includes additional information about the date the loan was issued and the approximate address (state and leading zip code digits) of the applicant.

Key Data Analysis Questions

With regards to the credit risk model you will develop, management has a number of concerns:

1. How does your new model perform compared to the one you used previously? How can it be expected to perform on new loan applications? For this, you must use the training and validation data used in the previous benchmark (Figure 1).
2. What are the important variables in this model and how do they compare to variables that are traditionally important for predicting credit risk in the banking sector? One regulatory requirement for lenders is that they need to to clearly explain how a loan application was assessed. To demonstrate to management that this can be achieved, clearly interpret all covariates in your model in terms of their effect on predicting credit risk.

Management has been talking to an expert consultant in statistical analysis (Prof. A. N. Omynous) about your credit risk modelling task. Prof. Omynous has suggested that one should be careful to account for variation in trends that may exist over time or between different jurisdictions. Management has never considered this before, so they have also asked you to use the *extended version* of the data to address the questions:

3. Can accounting for this variation (e.g., state/zip-code and time) improve performance benchmarks?
4. Are there any surprising differences in variables that are important for predicting credit risk? This is again essential for regulations.
5. Does credit risk change over time or between states? This is not something the bank has previously investigated and results may inform modified loan policies in the future.

Assignment Tasks

Project Presentation

You are required to present your models and findings to management. Even though your analysis may not be complete at this stage, you will still need to present models that address the objectives, concerns and questions outlined above. The presentation needs to have slides and they need to be engaging. Expect to answer questions about any aspect of the analysis or your presentation. Management may give you feedback on your model or modelling process that you can incorporate into your final report.

The presentation will run in person during the Monday Lecture and Practical timeslots in Week 12, see the task description for further details. You are not required to submit your slides. Otherwise, consult the criteria document on Blackboard to see how the presentation will be marked.

Analysis Report

Your team will conduct a regression analysis using generalised linear models and generalised linear mixed effects models. The response variable needs to be the loan default variable. This is a binary response (1 = loan default, 0 = loan repaid). You will then write an analysis report that addresses the above objectives and management concerns.

This is your team's final analysis report on your modelling which you will motivate using the above objectives, management concerns and questions. Document and develop your analysis in a single Rmarkdown document. The audience of this document is another data analyst.
There are two models you need to develop, a generalised linear model to address questions 1) and 2), and a generalised linear mixed effects model to address questions 3)–5).

In all cases, you must validate the assumptions of any models (when possible), and assess the performance of the model using industry applicable methods such as cross validation. For variable selection you **may** consider differrnce of deviance tests (likelihood ratio tests) or infomation criteria (AIC or BIC) , but you **must** consider the Gini score as one of your preferred measures for model selection.

You are required to submit the RMarkdown and PDF files via Blackboard. You are also required to submit the data set for reproducibility.

Summary on a Page (SOAP)

Produce a 1–2 page summary for management that addresses the above objectives, management concerns, and questions. This must include at least one plot. Utilise graphics creatively to make your points clear wherever possible. Some considerations:

- Nominate the methods used but do not describe them in detail.
- Base your assertions and recommendations on evidence from your analysis.
- Present both important variables and their effects in addition to some analysis of variables your management group may have been expecting to be important (if not present in final model).
- Do not present the effect of a covariate without communicating the uncertainty around that effect. State confidence intervals and show confidence bounds on plots.
- Be concise. Dot points are appropriate.
- This is not the work, it is like the advertisement for your work. In the real world, people are unlikely to look at the work if the advertisement isn't clear and engaging.

This is to be submitted with your group report as a PDF document. You are encouraged to use Rmarkdown for this document, however, this is not a requirement.

Individual Performance Review

This is a review that contains a short list of questions that will give you the opportunity to discuss your personal contribution to your team project (to be verified by each member of your team). Each team member will submit their own Performance Review.

Submission

Submission of this assessment with be electronic via Blackboard. Please note that this assessment item is due at 11.59pm Friday, Week 13, 2022. This is a strict deadline, and only files submitted by the time will be marked. Hence, it is worth submitting your assignment early and double checking that you have attached the correct file.

Submission Format

Please submit a single compressed .zip file. Keeping your submission neat and tidy will assist in grading. Create a README.txt file if you need to give me some instructions. Ideally your repository will contain only:

1. Your analysis in Rmarkdown form. e.g. `Credit_Risk_2022.Rmd`
2. The data files that you got from Blackboard.
3. Your SOAP e.g. `Credit_Risk_SOAP_2022.Rmd` or `Credit_Risj_SOAP_2022.pdf` etc.
4. Your README.txt file. (Optional)

Academic Integrity

The tasks in this project are to be completed as a *group*. It is expected that each member of the team contributes a fair shared of the workload, and has the opportunity to do so. Your groups submission must be your own work. You are not permitted to copy, summarise, or paraphrase the work of other groups in you submission.