Project: Predicting Loan Defaults with Logistic Regression

DS705

Overview

Use logistic regression to predict which applicants are likely to default on their loans. The dataset along with descriptions of the all of the variables can be found here:

https://datascienceuwl.github.io/Project2018/TheData.html (https://datascienceuwl.github.io/Project2018/TheData.html)

You're going to build one report progressively in three stages: Part 1, Part2, and Final. Parts 1 and 2 serve as checkpoints so that we can give you feedback to help you achieve good results with your Final submission. At each stage you'll submit both an .Rmd file and a knitted Word Document. Detailed instructions for each stage are given in additional sections below. Please visit the course calendar for the due dates associated with the three stages of your report. The final report **should be no more than 15 knitted pages.**

You'll be graded on two elements. The first element is having correct technical content and thorough analysis that addresses the question prompts. At each stage you must have working R code and a knittable .Rmd file. The second element is having an appropriate and well-written narrative throughout the report. Your narrative should explain what you are doing and why you are doing it. Your narrative should respond to the question prompts at each state below. The narrative should be written so that **someone with a similar statistical background**, but no knowledge of this particular project, **could read along and reconstruct your analysis**. You should choose graphs and tables that support your conclusions. Grammar, spelling, and clarity count for the narrative.

Effective communication means sharing enough but not too much. You shouldn't share all of the output or even all of the code in your final report. For instance, if you present a histogram you don't need to show the code, just show the well-labeled result and refer to it in your narrative. Showing a really long model summary with lots of p-values probably isn't useful and could be better summarized in a sentence or two. You can add options to your R chunk headers to hide unnecessary details, change the dimensions of figures, or even to suppress evaluation of a chunk of code that you just want to share. Add `echo=FALSE` to see the output but not the code. Add `eval=FALSE` to not evaluate the code in the chunk. Add `message=FALSE` to suppress unnecessary loading messages and such. More information about R chunk options can be found in the Rstudio -> Help -> Cheatsheets menu.

Unlike most homework problems, there isn't a single right answer in this project. There are many choices to make in building a predictive model and we'll be looking to see how well you've explained your choices as well as how you've analyzed the impact of those choices.

Format and Examples

Your final report (max 15 knitted pages including Executive Summary) should include the following sections in this order:

1. Executive Summary (max 1 page)
2. Introduction
3. Preparing and Cleaning the Data
4. Exploring and Transforming the Data
5. The Logistic Model
6. Optimizing the Threshold for Accuracy
7. Optimizing the Threshold for Profit
8. Results Summary

Details about the content of each of these sections are provided in the instructions sections later in this document.

Here is a good example of a logistic regression project, unrelated to this class, that will give you an idea of what we expect in your final report:

https://rpubs.com/mudassirkhan19/regression-model (https://rpubs.com/mudassirkhan19/regression-model).

This report includes a good narrative that guides the reader and highlights the results. This report would receive a grade of A in our class. However, it isn't really necessary to show exactly how every plot is produced unless the code is very unusual.

Here is an example of a report which would likely earn a grade of B:

https://rpubs.com/cheverne/titanic_survival_logisticreg_svm (https://rpubs.com/cheverne/titanic_survival_logisticreg_svm)

This report doesn't hide any of the unnecessary output and the narrative doesn't really serve to highlight the results and analysis. Instead the narrative just kind of tells us what the author did at each stage but doesn't explain why and doesn't focus on the highlights.

Points, Due Dates, and Grading

- Part 1: 25 points (15 technical, 10 narrative).
- Part 2: 25 points (15 technical, 10 narrative).
- Final: 50 points (20 technical, 15 overall narrative, 15 Executive Summary).

For Parts 1 and 2 we'll try, as always, to give you feedback within one week. For the final report we'll grade and give feedback before grades are due.

Instructions for Part 1

Complete Sections 2-4 and upload the Rmd and Word files to the D2L dropbox. You need to do all of the technical work described above and write a narrative explaining what you're doing and why you've made the choices you have. Refer to the examples above to see what kind of narrative we expect. Your Part 1 report should include section headers and content corresponding to Sections 2-4 as described below. Sections 3 and 4 are similar and can be combined or organized differently if you choose. You'll have opportunities to improve your work in the later stages, but you need to address the following elements in Part 1:

Instructions for Section 2 - "Introduction".

Write a brief statement describing the problem and how you will approach solving it. For everything except the Executive Summary your target audience is someone with statistical knowledge similar to yours who will be trying to reproduce your analysis. Think of this section as setting the stage for someone to reproduce your work.

Instructions for Section 3 - "Preparing and Cleaning the data".

Show how you've cleaned, prepared, and explored the data. Start by downloading the data through the link at the top of the document. Things to do in the section:

- Prepare the response variable based on the values of the **status** variable. Your response variable is a new version of the **status** variable and will be a factor variable that has two levels: "Good" and "Bad". Good loans are all those that are fully paid. Bad loans are loans that have a status of charged off or default (there may not be any "default" in this data). Loans that are late, current (being paid), or in grace period should be removed from the data.
- Eliminate variables that you think are clearly not useful as predictors and explain your choices.
- Do some feature engineering. For instance, some of the categorical variables (features) have levels with only a few cases. Consider lumping all of the small categories into one larger "other" category. If there are obvious redundancies among the variables, then explain and remove those variables. Creating new variables from existing ones is another possibility but is not required. As usual, explain your choices.
- Deal with missing values. How many cases now have missing values? Does it make sense to try to impute these values (in DS700 you may have used the 'mice' package to approximate missing values) or is it OK to delete the cases with missing values? Explain your choices.
- Explain what you are doing and why you are doing it. Look at the first example above to see what sort of writing we expect.

Instructions for Section 4 - "Exploring and Transforming the data"

- The modeling technique you'll be using in Part 2 is called Logistic Regression. There are no special requirements on the distributions of the predictor variables, but a predictor variable that is skewed is likely to have extreme values that can greatly influence the resulting model. Look at the distributions of the quantitative predictor variables. Are any of them strongly skewed? (Income related variables tend to have positive skew.) Identify the variables with very strong skew (income for instance, but the loan amounts are only mildly skewed) and replace those predictors in the data frame with transformed values. (Typical transformations for skewed data are square roots, cube roots, etc, logarithms, and reciprocals. Ideally the transformed values

would have fewer and less extreme values, and the distribution would be more normal in shape.) Explain which predictors you've transformed and why. A nice discussion of transformations may be found here: http://fmwww.bc.edu/repec/bocode/t/transint.html (http://fmwww.bc.edu/repec/bocode/t/transint.html)

- Every data analysis project should include some data exploration. In this case we're trying to predict loan status. Make graphs to explore the relationships between the predictors and loan status. For instance you could make a side-by-side boxplot of a quantitative variable to see if the variable is distributed differently for good and bad loans. Overlaid density curves, side-by-side or interwoven histograms can be used similarly. For categorical predictors you can use bar graphs or tables to show how the category distribution varies for good and bad loans. Are there any predictors which behave quite differently for good and bad loans?

Instructions for Part 2

Work on improving Sections 2-4, if necessary, in response to feedback you may have received. Complete Sections 5-8 and upload the Rmd and Word files (including all sections so far) to the D2L dropbox. Your Part 2 submission needs to include all of Sections 2-8 and Sections 2-4 should incorporate improvements suggested by you instructor in response to Part 1. You'll have opportunities to improve your work in the final stage, but you need to address the following elements in Part 2:

Instructions for Section 5 - "The Logistic Model".

This is your first attempt at a model.

- Create two datasets from your cleaned and prepared data. Randomly choose 80% of the cases and make this into a "training" dataset that will be used to build your logistic regression models. You may wish to add a `set.seed(integer of your choice)` before you randomly choose cases so that your results will be consistent each time you run your script or knit. The remaining 20% of the cases are your "test" or "validation" dataset. You'll use the test dataset along with your model and `predict()` to generated predicted statuses for each loan and to analyze the performance (accuracy) of your model.
- Create a logistic regression model, using the training data, that uses all of your remaining predictors to predict loan status. Remember you can't use **totalPaid** as a predictor so you may wish to remove that column from the training data. However, keep the **totalPaid** column in the test data because you'll want to use it later.
- Use your model to predict the loan status for loans in the test data set. Use the default threshold of 0.5 to classify the "Good" and "Bad" loans. (Don't forget to use predict() with type="response" to estimate the probabilities.) Use the predicted loan status to make a contingency table and also determine the overall accuracy of the model (the percentage of correctly predicted outcomes), the percentage of actually good loans that are predicted as good, and the percentage of actually bad loans that are predicted as bad.
- Work on the narrative for this section. Is this an effective model for predicting if a loan will be repaid?

Instructions for Section 6 - "Optimizing the Threshold for Accuracy"

By varying the classification threshold from 0.5 you can correctly predict more bad loans. Throughout this section investigate how changing the threshold affects your model predictions when applied to the test data.

- Experiment with the classification threshold to change the proportions of loans that are predicted as good and bad. As the threshold varies from 0 to 1 what is the effect on overall accuracy and proportions of correctly predicted good and bad loans? A perfect way to communicate this information is with a graph showing accuracies versus threshold. What value of the threshold produces maximum overall accuracy and what is that accuracy? Explore the tradeoff between correctly predicting good and bad loans.

Instructions for Section 7 - "Optimizing the Threshold for Profit"

From the bank's perspective the most important feature of the model is how it changes the overall profit. Repeat the threshold analysis of the previous section, but this time find the value of the threshold that maximizes the profit. Compute the profit by applying your model to the test data and assume the bank denies all of the loans that your model predicts as "bad". How does this change the total profit? (Hint: for the loans that are predicted as good you need to find the sum of **totalPaid - amount**.) Now investigate how changing the classification threshold affects the total profit if loans that are predicted as bad are denied by the bank. Compared to not using your model, what is the maximum percentage increase in profit that can be expected by deploying your model? How does this increase in

profit compare to the increase in profit from a perfect model that denies all of the truly bad loans? For your best profit threshold, what is the overall accuracy and percentages of correctly predicted good and bad loans? Does the maximum profit threshold coincide with the maximum accuracy threshold?

Instructions for Section 8 - “Results Summary”

Here is where you will summarize a few details for the final classification model you will suggest for the bank. Be sure to include your final value of the classification threshold as well as a summary of the overall profit and accuracy of the model including a breakdown of the percentages of correctly predicted good and bad loans. Your goal in this section is summarize the final model and its effectiveness for someone trying to reproduce your analysis. This section is different than the Executive Summary in the Final Report where you will summarize the benefits of the model for management.

Instructions for Final Report

Incorporate feedback you may have received about Sections 2-8 to improve your final report. Write the Executive Summary. The Executive Summary will be the first page of your report and should explain what you’ve done and why the bank should adopt your model to increase profit. The Executive Summary should be seen as a stand alone document that will be read by the top management of the bank. Keep in mind that management **does not know statistics**. You’ve got one page to briefly explain your methodology, validation, and reasons why the bank should deploy your model. A graph would be OK in the executive summary if isn’t overly technical and easily communicates the main story about increased profit. How will the model change both profit and the number of loans awarded by the bank? Your goal here is to provide management with enough easy-to-read information to be able to decide whether or not to deploy your model. This is not an introduction, instead it is a one-page summary written for a different audience than the rest of the report.

Your Executive Summary will graded based on both how well you summarize the project and how well you address the appropriate audience.

This site gives a couple of examples of a good summary and one example of a bad summary:

https://unilearning.uow.edu.au/report/4bi1.html (https://unilearning.uow.edu.au/report/4bi1.html)

The final report should be no more than 15 pages including the Executive Summary. Upload both the Rmd and Word files to the dropbox.