Big Data Applications in Finance
Individual Assignment

Daniele Bianchi

September 18, 2021

Details

Grading: 20% of your final grade.

Deadline: 20 April. Time: 22:00hrs (London time).

Submission mode: QMplus

Submission files: 1) An assignment_results.m file with the code, and 2) A "Prediction Report" with maximum of 5-pages with the interpretation of your results.

Please refrain from copying code from each other. Your code and output submission will provide us sufficient information to detect such practices and it will be penalized.

What am I grading you on?

I want to test your ability to use models learned during the first part of the course and to synthesize the insights in a concise way into a report. This assignment requires you to apply all the skills learned during your first part of the term, and the models learned during the lectures. Your ability to estimate the model is only half of the challenge. Your interpretation of the findings matter equally.

Can I ask assignment related questions during office hours?

The questions that your TA and I will answer are only to clarify the meaning of a question. We will not provide any clue on whether what you are doing is "right", or "wrong", or suggest ways to improve your code. However, if you have doubts on estimating a model, we then ask you to phrase your questions on the tutorial datasets and models, so that we are able to provide guidance.

The dataset provided to you should not have many challenges to resolve before you estimate a model. However, if you face any challenges while using the data on Matlab, do reach out to us. I strongly recommend that you work on your assignment in your computer, and not on the cloud so that you have complete control over the file. Any excuses that the file was not "saved", or went "missing" on the cloud will not be entertained.

Submission files

You are expected to create a .mat file with the results and an .m file with the code as part of your submission. The .m file should be self-contained, i.e., I should be able to run the code on my computer loading the same data file without any error and ideally obtain what there is in the .mat file. In addition to the Matlab code, you are expected to create a PDF submission of no more than 5 pages, that present your evaluation of the models.

The Assignment

LendingClub is an American peer-to-peer lending company, currently the world's largest platform that allows for individuals to both invest and borrow on the platform. Borrowers can obtain unsecured personal loans from the platform, and this assignment is set up for you to assess your ability to predict defaulters in the data using the predictors provided in the data.

The data is a random sample of loans issued on the platform between $2007 - 2015$, including the loan status, and payment information. The data also contains a number of predictors that have been documented in the variables description file provided to you named "ECOM151-Assignment-VariableDescription.xlsx". For tractability, your assignment focuses only on a small set of variables

available for prediction.

You have been provided with on three .csv data files:

trainData: This is the dataset on which you will train all your models.

testData: This is the dataset on which you will evaluate your model's fit.

varDescription: This is a replication of the variable description available in the excel spreadsheet provided to you.

Question A (10 points)

This question expects you to estimate five different class of models to identify the best model to predict default on the LendingClub platform.

Set up the data

Load the files as Tables on your work environment.

Questions:

1. Create a new variable from trainData called "y" which takes the value = 1 if loan_status is "Charged Off" and 0 otherwise.

2. All variables provided to you other than loan_status are referred to as "predictors".

3. Find the 10 most correlated variables with y.

4. Find the 10 lowest correlated variables with y.

Now, we are ready to run the five models of interest. Spend time, and visualize the data. Inspect for potential reasons why the model may not be estimated.

Pay particular attention to whether you would like to transform your variables (For example, a logarithmic transformation). This may also help with interpretting coefficients in your "Prediction Report". You may also want to consider converting some of the categorical variables into a continuous variable.

6. LINEAR REGRESSION MODEL: Fit a linear regression model to the trainData, with y as the outcome variable, with all the predictors.

   (a) What is the Mean squared error for the training data?

   (b) What is the Mean squared error for the testing data?

7. BEST SUBSET MODEL: Fit a "subset selection" model to the trainData, with y as the outcome variable using the stepwise regression approach. Hint: the command you should use is *stepwiseglm* in Matlab (see link `https://www.mathworks.com/help/stats/stepwiseglm.html`). We will cover this in the tutorial so you should use the same estimation framework. Please explain is you use different estimation assumptions you take.

   (a) What is the Mean squared error for the "best" model of this class for the training data?

   (b) What is the Mean squared error for the "best" model of this class for the test data?

   (c) What are the variables in the "best" model of this class?

8. RIDGE REGRESSION MODEL: Fit a ridge regression model to the trainData, with y as the outcome variable, with the predictors. Hint: the command you should use is *ridge* in Matlab (see this `https://www.mathworks.com/help/stats/ridge.html`). We will cover this in the tutorial so you should use the same estimation framework. Please explain is you use different estimation assumptions you take.

   Explore all values of lambda (10^{10} to 10^{-2}), setting lambda = 0 initially. This should be roughly equivalent to the OLS estimation.

   (a) What is the Mean squared error for the "best" model of this class for the training data?

   (b) What is the Mean squared error for the "best" model of this class for the test data?

   (c) What are the 10 most important variables in the "best" model of this class?

9. LASSO: Fit a LASSO to the trainData, with y as the outcome variable, with the predictors. Hint: the command you should use is *lasso* in Matlab (see this `https://www.mathworks.com/help/stats/lasso.html`). Explore the same values of lambda as for the ridge regression. We will cover this in the tutorial so you should use the same estimation framework. Please explain is you use different estimation assumptions you take.

    (a) What is the Mean squared error for the "best" model of this class for the training data?

    (b) What is the Mean squared error for the "best" model of this class for the test data?

    (c) What are the 10 most important variables in the "best" model of this class?

10. RANDOMFOREST: Fit a randomForest to the trainData, with y as the outcome variable, with the predictors. Explore and fit the best model of this class. Hint: the command you should use is *fitrensemble* in Matlab (see this `https://www.mathworks.com/help/stats/fitrensemble.html`). We will cover this in the tutorial so you should use the same estimation framework. Please explain is you use different estimation assumptions you take.

    (a) What is the Mean squared error for the "best" model of this class for the training data?

    (b) What is the Mean squared error for the "best" model of this class for the test data?

    (c) How important are the variables in predicting default?

11. Compare and contrast the predictive power of all approaches and identify the best model to predict default in the LendingClub data.

Question B (10 points)

You are required to synthesize all the work in Question A to submit a "Prediction Report" to your manager on your ability to predict default for borrowers on the LendingClub platform. Utilize all the information you have generated to write a report no longer than 5 pages and present your best model to your manager. Pay attention to explain why it is the best model, in terms of its out of

sample predictive power, and visualize the model's predictive power compared to the other models on hand.