Sierra Moen
Spring 2017
Math 216

Project: Give Me Some Credit
Description: Improve on the state of the art in credit scoring by predicting the probability that somebody will experience financial distress in the next two years.

My goals: Load data and do EDA
Find optimal lambda for Ridge Regression and LASSO, compare between methods
Create an ROC curve to make more accurate model!

5/9 1.5 hours

I chose to enter the 'Give Me Some Credit' competition so I can build off of the code produced in PS08. I plan to look over 2-3 of my classmates' codes to use as a starting point. I will improve upon them by incorporating other aspects of machine learning used in this class.

I first looked over Aayam's and Rebecca C's Kaggle entries to get an idea of what they did. I chose them because they don't look too in depth, so I will have ample opportunities to improve the model (although I think they did a great job!)

First, I'll run both of those codes to see how they work!
Aayam's code looked good to me! There were a lot of things I did not understand, like some of the log/probability stuff. The Kaggle score on Aayam's submission was .587. The leading score was .870.
I was unable to run RC's code, which was very frustrating! I got A TON of error messages so I decided to wait and ask for help as to why that was the case.

Second, I'll do EDA on the variables in the data. I skimmed some of the other Github files and saw Pastora plotted each variable - I'll do that but using my own skills!

5/11 1.5 hours

I am ready to make my own forays with this data! I want to clean the data and do an exhaustive EDA.

While renaming the variables: THIS is the type of frustrating stuff I was talking about! I have spent nearly 10 minutes trying and failing to rename two variables, while the other variables rename easily. I would just leave it for now, but I have to rename these because R won't allow '-' symbols. It seems like I should note that this part of the exercise has caused physical pain.

I get the message : Error in rename_(.data, .dots = lazyeval::lazy_dots(...)) : argument ".data" is missing, with no default

Oh I am mad now because it worked for no reason – I legit didn't change anything and the code ran. I wish I had been able to spend the past 30 minutes doing actual work on this assignment. I just spent another 20 minutes trying to code how to replace NA's with the mean of the column. Yes, I am learning to 'work through my coding problems,' but, believe me, at this point I have so many real world problems to work on that this hardly seems like a good use of time. Plus, when I code it really does produce physical pain and makes me cry (only partly due to the pain). I'm not willing to compromise my health for this project any more then I already have so this is a pickle.

5/14/17 2.5 hours

I have decided to take a different approach to enacting my project. Since I had SUCH a hard time changing the name of a variable, I will first write buckets of pseudo-code to demonstrate what I'm going to do on the project. That way, I can flex my data muscles without wasting so much time. After a 20 minutes, I decided to temporarily use Albert Pastora's EDA. I really did want to do my own analysis of the variables using different types of plots (bar and box), but had serious issues the last time I tried. He did something strange with $ and 'group,' so that is not something I would be able to replicate.

From my analysis of the graphs (variables are 'good' if the lines on the graph don't overlap very much): Revolving Utilization of Unsecured Lines, Number of times 30-59 days past due; Debt ratio; Number of times 90 days late, Number of times 60-89 days past due are variables that are useful predictors. If I have time in the future I would like to make my own EDA, but this was definitely a good choice for today!

I wrote out pseudo code to find optimal lambda using Ridge Regression and LASSO. I then wrote it out in R (borrowing from PS06). I can't run it yet because I need to have the variables cleaned up before I can plot them, and I need help to do so. I then prepared to find the best sensitivity using ROC curves, borrowing from PS07. I actually referred to my *own* homework – yay! I didn't run it though, because I both want to do LASSO and RR first and I'm not sure what p_hat will end up being.

5/15/17 2 hours

Goals for today: Get data cleaned! Create own EDA for at least three variables. Run the LASSO and RR code.

Totally figured out the is.na(train) thing – no more N/A's! Prof. showed code to get mean.

EDA: About 100 people have VERY late bills, which is skewing my data! Let's see if I can figure out how to skim them off! Yes I did it!

Ridge Regression: I ran RR code and it went pretty smoothly! I got a little confused about why my graph wasn't the colorful one, and the answer was that I hadn't run that chunk of code yet!

5/16 3 hours

Goals for today: finish RR and LASSO, begin ROC analysis.

RR & LASSO: I am doing this to decrease the variance of predictors, which has the effect of increasing bias. I ran the model several times. It showed that lm() was the method with the least MSE. I don't really believe that – there is definitely something wrong with my code. Unfortunately, I don't have the technical skills to fix this and was unaware there would be no office hours after Tuesday. I wish I had known that earlier then yesterday, as I won't be able to comprehend coding or analysis advice over Slack. As such, this 'journal' will be longer then 1 page so I can adequately display the depth of my analysis.

ROC: I made my ROC code using logistic_regression as a starter code. It went pretty well in terms of me figuring stuff out without spending that much time on little details.

.695 .704 with all variables. Score decreases when variables are removed. This makes sense but I'm sad I can't apply my EDA analysis.

5/19 3 hours

Goals for today: finish analysis, create files to submit, submit files!

I am stuck on my first goal! Ran into a lot of issues when finishing analysis and trying to submit entry to Kaggle. I will have to wait to go to office hours.

5/21/17 3.5 hours

Goals for today: Same as 5/19!

Since I'm predicting a binary variable, I can't use lm, which was what my LASSO/Ridge Regression/lm analysis told me to do. My final model seems pretty like a cop out, since I'm just using all the variables and not turning many 'knobs.' However, according to all my EDA, glm was indeed the best model to use!