Directions for Final Project Spring 2019.

On Monday, April 29th at 9:15-11:15 a.m. will be final presentations. Guidelines/rules:

1. **Work individually**.  This is your creation.  This is different than previous assignments. It is perfectly acceptable to review all student 'keys' posted over the course of the semester and use code and ideas from those projects.
2. You will present individually.  You should plan for 15 minutes with max 15 slides.
3. You need to submit at least 3 items.  a) your code file of the code you used on the project.  b) A 'white paper' written report in either pdf or Word document form. c) a Powerpoint presentation.  If your document ‘white paper’ was created with R Markdown, submit both the document and the .rmd file.

Background and general guidance: Well our executive group was shocked we had so many ‘bad’ customers, but also taken aback a bit by how many really profitable (good customers) there are and really is wondering what we should do about that. Well Rex pointed out (Rex is a real person by the way….you should request to be linkedIn with him on LinkedIn Rex Pruitt….very knowledgeable and influential). Anyway Rex pointed out that he already made such a model and called it ‘Good Customer Score’ (true story). Now we don’t know exactly what Rex used as predictors but we do know a few things and we think we can ‘beat Rex’. We know that Rex used to work for an insurance company and there are many similarities between risk weighting insurance factors and bad credit. Car crashes and bad loans aren’t that different. We know this score was meant to be separate from other ‘bureau scores’ such as FICO and Behavior Score. And we know that this was about finding the good customers so that we can do something as incentive to make the customers use our card even more and make more money. So we want to:

1. Focus only on those customers that we will be likely to give something to as incentive. These might include credit line increases or annual fee waivers, etc. Stuff that costs money. Think about it. If you give 1 million customers a $500 credit line increase that’s $500,000,000….you better be right.
2. NOT use Behavior Score or FICO in the analysis.

Instead we want to make a model (or more than one ‘candidate’ model) from the data we already have from the midterm with the following constraints…..

1. Remove any customers from the data who have as Row Number =1 that has either a Days Past Due > 0, or a non-blank External Status, or who have an ending or beginning balance that month which is > Credit Limit. We found out from the midterm that those particular customers are just too risky and we aren’t about to give them more money…we’ll just lose it.   So again if any of those conditions are true, remove those customers.
2. We need to be a little stricter on our definitions of good and bad. However we still want to just use a binary target of 0/1. NO OTHER types. But we are torn on the definitions. Some people argue by removing those in step 1 we already got rid of the bulk of the bad customers so we want to create a model that is about closed vs not closed (status = C) voluntarily. Others think no no…stick with good/bad being some kind of non-payment. So pick from these. They are TWO DIFFERENT KINDS OF BAD. There is still a little purposely ambiguous nature left in both choices.
   1. Bad = 1 defined as Last Known Status (final month) is 90 or more days past due OR any External Status other than blank or ‘C’. This status must be Month 7 or greater. Else Bad = 0.
   2. Bad = 1 defined as Last Known Status (final month) is External Status = ‘C’ and days past due = 0. This status must be Month 7 or greater. Else Bad = 0.
3. Another BIG change is which are allowed input variables.  DO NOT USE FICO OR BEHAVIOR SCORE OR GOOD CUSTOMER SCORE.  You will have to scrutinize each of the other variables more closely to find any meaningful variables.
4. Many people glossed over problematic variables, didn’t read or follow directions on creation of variables, ignored missing data, ignored odd values, or otherwise loosely defined the issue. You will find that this becomes more challenging when removing the big predictor variables such as FICO or Behavior Score. Look at the previous directions again as well as what others did. Does your information make sense?
5. You must make at least 2 'candidate' models.  One should be logistic regression.  The other should be MARS which is a good example of a ‘fancy’ model. However occasionally people think they learned something more interesting. Feel free to include an extra type of model you may have learned elsewhere but it’s not necessary and consider yourself cautioned.
6. For input variables you MUST ‘bin’ at least 2 input variables. We’ll cover this in class. You MUST also calculate new transformed variables similarly to the midterm. Predictor variables ONLY come from Row Num =1 or transformations of that row.
7. In the midterm many groups 'hand waved' variable selection.  This worked because FICO, Behavior Score, and Good Customer Score are powerful predictors.  You now are tasked with 'interrogating the data' to find predictive power.  Thoroughly examine each input variable for inclusion.  Be prepared to answer 'why isn't variable 3 in the model'? Descriptives.  Besides histograms and things of that nature, plots of bad rate on x axis and input variable on the y axis are extremely useful to examine the relationship between target and input variables one at a time.  The easiest way to do this for continuous variable is to bin them (even if you don't use binning in the modeling method) then plot the bad rate for each range. It gives you ‘directionality’.
8. For results do all the normal things I've shown you.  KS, ROC, Lift, gains table, etc.
9. Did you beat Rex? You should compare your model's performance to the Good Customer Score. You can do this by considering Good Customer Score as just another model (which it is) and so you can calculate lift, KS, gains, etc for just Good Customer Score and compare to your model.  Basically you have ‘points’ instead of ‘probability’ for Good Customer Score, but ROC/KS/Lift etc are measures of ‘rank order’ which is retained (invariant). Good Customer Score is simply a probability or log-odds that someone transformed into points. The ROCit library vignette shows an example of this type. A far simpler comparison would be to use a Gains table as a basis for the comparison.
10. Lastly: Are we making money? Review the midterm guidance on profit and calculate profitability in your gains table based on the definitions from the midterm.