Assignment 5.2: Building a Hazard Model for Consumer Bankruptcy

Submission Details

- Submit through T-square dropbox
- You have to submit ONLY
  - SAS program
  - Output in PDF format
  - You don't need to submit any datasets
  - Short and concise explanation (1-3 lines) for the specification used in the regressions (economic rationale for each variable and expected sign). For example, I chose *credit utilization rate* as one of the explanatory variables because high utilization rate implies that a consumer tend to take big risks and higher utilization rate is expected to result in a higher (+) likelihood of default.
  - Do not run the regressions first, see the sign and then input this sign in the writeup. Provide an economic explanation for why the variable should matter for default prediction and how (positive or negative) it would effect the default likelihood.

Assignment Tasks

1. Come up with a list of possible covariates that should matter for default prediction. Some possible sources are
    - Lecture on credit scoring
    - Data Dictionary
    - Lending Club
    - Prosper
    - Lend Academy
    - Nickel Steamroller
2. Compute these explanatory variables
3. Do a *in-sample* estimation and prediction. Follow these steps
    - Use the entire time period 2007-2015 for estimation
    - Run a PROC LOGISTIC model with DESCENDING option with bankruptcy as the LHS variable and the variables in step 1 as explanatory variables
    - Present the output and fit statistics for the model (output of proc logistic).
4. Do a *out-of-sample* prediction. Follow these steps
    - Divide the sample into in-sample estimation period (2007-2013) and out of sample forecasting period (2014-2015)
    - Estimate the model with 2007-2013 data

- Forecast default for 2014-2015 time period using the estimates from 2007-2013 time period and explanatory variable data from 2014-2015
- Rank the default probabilities into deciles (10 groups). Use PROC RANK
- Compute the number (and percentage of defaults) in each of the 10 groups during 2014-2015 time period.
- A good model is one that has the majority of defaults in decile 1 or 2 and very few in other deciles
- You can use other metrics like ROC curve etc., but you don't need to compute those measures

5. iterate steps 1-4 till you get a model with good out of sample performance

Data Extraction

- Download customer credit datasets and data dictionary from Tsquare - LoanStats3a.csv, LoanStats3b.csv, LoanStats3c.csv, LoanStats3d.csv, and LCDataDictionary.xlsx, which have already been modified for your convenience.

- You can use PROC IMPORT to transfer these datasets

```
proc import datafile=\url{"c:\LoanStats3a.csv"}
out=LoanStats3a
dbms = csv replace getnames = yes;
run;
```

- Convert two character variables

```
\verb|mths_since_last_record| \\
\verb|mths_since_last_major_derog|
```

  to numeric variables in the LoanStats3a dataset. And then set these four dataset together to create a total sample set.

SAS commands that may be useful for the Data Analysis

As I mentioned in the class, SAS provides multiple methods to perform any given task. Some of the following commands may be useful (look up the examples from the UCLA SAS web site link that I posted on Tsquare)

- PROC LOGISTIC
- PROC MODEL
- PROC RANK
- PROC SQL
- PROC SORT
- PROC FREQ
- PROC UNIVARIATE
- PROC MEANS
- PROC PRINT
- PROC CORR

Data Analysis

Steps in the assignment

1. First, make a list of variables you expect to matter for default prediction.

2. Pre-process LoanStats data to create variables that you require. There are many important variables in the string format. Think about how to transform them into meaningful numeric variables. Some of them can be simply transformed into binary variables, like term (36 months or 60 months), while some of them can not, like employment length.

3. There are two variables - *fyear* and *label* need to be computed to segment data and train the model

```
if length(issue\_d) = 6 then \\
fyear = ("20" $ || $ substr(issue\_d,1,2)) + 0; \\
else fyear = ("200" $ || $ substr(issue\_d,1,1)) + 0;
```

4. Compute the required variables.

5. Compute the required statistics.

6. Use PROC LOGISTIC to model default occurrence.

7. Save the results in PDF format. I don't want to see hundreds of pages of output.

8. Check the LOG file to see if there are any errors in your code.

9. Upload the SAS program and results to tsquare.