# Data Analytics and Machine Learning | Prof. Lochstoer

## Problem Set 3

Use the LendingClub_LoanStats3a_v12.dta Stata dataset available at CCLE (Week 3) for this exercise. The data is downloaded from Lending Club's website. Feel free to have a look at the raw data and the variable description available there at https://www.lendingclub.com/info/download-data.action.

## Question 1: Predicting default

a. We will use the column "loan_status" as the indicator for whether the loan was paid or there was a default.

(i) Drop all rows where "loan_status" is not equal to either "Fully Paid" or "Charged Off." Define the new variable Default as 1 (or TRUE) if "loan_status" is equal to "Charged Off", and 0 (or FALSE) otherwise.

(ii) Report the average default rate in the sample (number of defaults divided by total number of loans)

b. LendingClub gives a "grade" to each borrower, designed as a score of each borrowers creditworthiness. The best grade is "A", the worst grade is "G".

(i) Using the glm function, run a logistic regression of the Default variable on the grade. Report and explain the regression output. I.e., what is the interpretation of the coefficients? Do the numbers 'make sense'.

(ii) Construct and report a test of whether the model performs better than the null model where only "beta0", and no conditioning information, is present in the logistic model.

(iii) Construct the lift table and the ROC curve for this model. Explain the interpretion of the numbers in the lift table and the lines and axis in the ROC curve. Does the model perform better than a random guess?

(iv) Assume that each loan is for $100, and that you make a $1 profit if there is no default, but lose $10 if there is a default (both given in present value terms to keep things easy). Using data from the ROC curve (True Positive Rate and False Positive Rate) along with the average rate of default (total number of defaults divided by total number of loans), what is the cutoff default probability you should use as your decision criterion to maximize profits? Plot the corresponding point on the ROC curve.

c. Next, we will see if it is possible to do better than the internal "grade"-variable, using other information about the borrower and the loan as provided by LendingClub.

(i) First, consider a logistic regression model that uses only loan amount (loan_amnt) and annual income (annual_inc) as explantory variables. Report the regression results.

Show the lift table, comparing to the 'grade'-model from a. Plot the ROC curves of both the 'grade'-model and the altnerative model. Which model performs better?

(ii) Now, include also information from the loan itself. In particular, include the maturity of the loan (term) and the interest rate (int_rate) in the logistic regression. Report the output. How does R handle the term-variable? In particular, what is the interpretation of the regression coefficient? Again show the lift table and ROC curve relative to the original 'grade' model. Now, which model is better? What is the likely explanation for why this new model performs better/worse?

(iii) Create the squared of the interest rate and add this variable to the last model. Is the coefficient on this variable significant? Please give an intuition for what the coefficients on both int_rate and its squared value imply for the relationship between defaults and the interest rate.