Project Data Description, Research, and Methodology

Joshua D. Ingram

New College of Florida

**LendingClub and Project Motivation**

LendingClub is a peer-to-peer lending platform that enables borrowers to have quick access to funds of up to $40,000, while lenders are able to assess each individual loan. This enables lenders to create a diversified portfolio that is in line with their risk level with potential for great returns. Like any investment, there are risks associated with investing through LendingClub. LenndingClub assigns each loan a grade based on their risk of default, from A to G, alphabetically. A loan with a grade A is the least risky, whereas as a loan with a grade G has the greatest risk default. The riskier the loan, the higher the interest rate and the higher the return for an investor if the borrower does not default. Assessment

Predicting the probability of default would be interesting with the data available, but considering loan grades partially fulfill this role, looking at how loan grades are determined by LendingClub could give valuable insight for potential lenders through the platform. Without direct access to the formula used, investors are left to trust LendingClub. Being able to understand what exactly may be used by the platform to grade the loans could allow investors to have even more useful information. Individual investors on this platform could make hundreds of decisions a day on whether or not to fund a loan, so understanding some of the variables associated with the loan grade granted by LendingClub grants the ability to save time and to make more informed decisions.

**Data Description**

The dataset being used is accessed through Kaggle.com, which was sourced from the LendingClub platform. The dataset contains data on all loans issued from 2007 through 2018, totaling 2.26 million observations and 145 variables. There is also an accompanying LendingClub data dictionary.

Dataset source: <https://www.kaggle.com/wendykan/lending-club-loan-data#loan.csv>

Due to the size of this dataset, below is a list of the most notable variables (although some may be dropped or more used):

Response

grade – LendingClub assigned loan grade (A, B, C, D, E, F, G)

Explanatory

*annualinc* – the self-reported annual income provided by the borrower during registration

*application\_type* – indicates whether the loan is an individual application or a joint application with two co-borrowers

*av\_cur\_bal* – average current balance of all accounts

*dti* – (debt-to-income ratio) a ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LendingClub loan, divided by the borrower’s self-reported income

*earliestCrLine* – the date the borrower’s earliest reported credit line was opened

*empLength* – employment length in years. (between 0 and 10. 0 – less than 1 year, 10 – 10 or more years)

*homeOwnership* – the home ownership status provided by the borrower during registration. (RENT, OWN, MORTGAGE, OTHER)

*isIncV* – indicates if income was verified by LendingClub, not verified, or if the income source was verified

*loanAmnt* – the listed amount of the loan applied for by the borrower. If at some point in time the credit department reduces the loan amount, then it will be reflected in this value

*mortAcc* – number of mortgage accounts

Due to the size of this dataset and limited resources, there is consideration of the removal of a number of rows, so the data is more manageable. The criteria for removing the rows would be dependent upon the time (e.g. loans from 2007-20011 may be removed) and/or whether the loan was requested by more than 1 borrower. This is still under consideration and would be stated in the final portion of this project.

**Project Methodology and Comments**

The focus of this project will be to properly utilize the models and practices learned in Introduction to Categorical Data Analysis. Since our response, loan grade, is multinomial and we are interested in interpretation, the multinomial logistic regression model will be used for this project. The model will be simpler, containing a small selection of predictors, so that I can understand and explain the effects of different variables on the probability of a loan being given a certain grade. If time permits, it would be interesting to create a model focuses on predictability. First, I would split the data up into training and testing data, then perform model comparison to select the model that has the lowest test error. I will likely use a random forest model, but that will be determined in analysis later.

The main concerns I have with this project at the moment is with its size, missing values, and differentiation and separation between joint and single loan applications. These will be attempted to be solved during the project and any changes to the dataset will, of course, be explicitly stated.

**Graphical and Numerical Data Summaries**