## Modeling Challenge

Introduction

In this case study, we are asking you to review a dataset, clean the dataset, and then provide us with your thoughts regarding 1) which business rules can be used to reduce default rate, and 2) how a model could be built to effectively predict a potential borrower’s chance of default.

Section 1: Data Review and Dependent Variable Definition

Please review the sample dataset contained in “data.csv”.

The sample dataset contains information about fictional loans that were issued between January 2015 and September 2016. To help you navigate this dataset, below is a short list of variable definitions:

* “issue\_d” records the issuance month of each loan
* "loan\_status" records the latest status as of 01/23/2017
  + "Current": The borrower has paid off all due payment as of the latest due date.
  + "Fully Paid": The borrower has paid off the entire balance of the loan.
  + "Default": The borrower has missed the last payment.

Additional variable definitions are contained in “data\_dictionary.csv”.

First task is to define an appropriate dependent variable. Please indicate if you will treat it as binary or multi-class classification challenge, and provide reasons for your choice. Please remember, the ultimate goal is to predict a potential borrower’s chance of default. Also, when making your determination, please consider if all defaults are created equal.

* The most useful dependent variable is probably **loan status**, since it indicates default. However, since it only indicates the CURRENT status of the loan, it may miss previous delinquencies on this same loan.
* I also considered using “**duration**” which is a common way to measure both term length and amount of a loan, but not enough information was available to construct this.
* For simplicity we will consider all defaults equal. If we were able to use the duration variable mentioned above, then we could theoretically gauge the severity of the default.

Section 2: Data Cleaning

Now that the dependent variable has been identified, please leverage Python to clean the sample dataset. Your goal is to get the data into a state where it can be fed into a model. For example, the variable, "earliest\_cr\_line" was recorded in the form of month-year and the year was a mix of the last two digits and the full four digits. This would need to be standardized before converting it into a numeric variable. In addition to data format issues, there might be a few variables that can cause [data leakage](https://www.kaggle.com/wiki/Leakage).

In the document where you defined the dependent variable, please briefly discuss your data cleaning methodology and findings. Please also attach a copy of the cleaned data and the code used to clean the data with your submission.

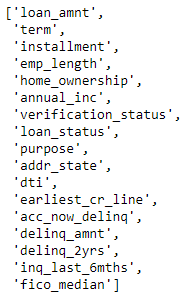
Section 3: Analysis

First, please try to answer the following questions:

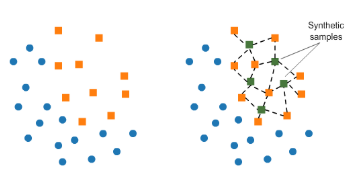
* What variables, if any, can NOT be used to predict a potential borrower’s chance of default? For example, information that happens after the loan underwriting decision is made.
  + I dropped the following columns because more than 1/3rd of the datapoints were NaNs: mths\_since\_last\_delinq, mths\_since\_last\_record, inq\_last\_12m
  + I dropped customer “ID” because it is an arbitrary and unique identifier
  + I dropped “issue\_d” because all loans were issued within a relatively narrow timeframe (Jan ’15-Sep ’16) and without a longer time period it wouldn’t be advised to be making a time series
  + I dropped the following variables because they would not be available by the lending decision: last\_FICO, last\_credit\_pull\_d
* Are there any ways to derive additional variables that would improve the model prediction accuracy?
  + I used *median* FICO scores instead of high & low because the two were highly correlated. I also considered including the “distance” spread between high and low as an additional variable, but since they almost always had a value of “4” I did not include it.
  + I converted earliest\_cr\_line from a date to “age.” The age was measured from the reporting date of the data (mentioned above) as 1/23/2017. Using an “age” variable better quantifies time for modeling instead of using dates.
* What variables, if any, can be used to create business rules that can be used to decline customer’s application before the model runs?
  + To determine if there were any simple “rules of thumb” that can be used to pre-screen customers, I ran an R^2 measure of Loan Status (having 0 for non-default, 1 for default) to every other variable that would be available before approving the loan.
  + Unfortunately, all the pairwise R^2 measures were very low, indicating there is no simple rule of thumb that can be used to prescreen. As a result, if default is in fact predictable, it likely needs to consider multiple variables via a model.

Then, please try to build a model to predict a potential borrower’s chance of default, and please briefly describe your model strategy including:

* Your choice of the classification method that you believe would best perform with this dataset.
  + I decided to start simple with a logistic regression model, and a simple classification tree as a nonlinear candidate. Both are widely used and powerful models for classification.
    - Trees can be further enhanced using boosting and bagging.
    - Grid search can be used to search a variety of hyperparameter combinations
* Any additional data challenges you may face given your choice of modeling methodology.
  + In order for the models to work some variables needed NAs to be imputed. I used random selection since the proportion was relatively small (<5%).
  + Since all of my models predict probabilities of default, we can choose different thresholds of probability cut-off to determine when we predict a default and thus decline a customer. However, it’s a trade-off -- the more people you reject then the less revenue you earn. In order to decide an optimal threshold, I would need more information on the revenue of non-defaulter vs. cost of default.
* The final list of variables that would go into your model after performing any variable selection technique that you deem necessary.
  + This was the final list of variables before I codified them:



* How you would validate the model.
  + I split the samples into train and test, while “stratified” by loan\_status to preserve the proportions of default.
  + Because the dataset was imbalanced, I used SMOTE method to synthetically boost the proportion of default samples in the training set. SMOTE is more useful than random-sampling-with-replacement because it creates “new” data that is an average of local neighbor datapoints. This allows for some extra variability that likely represents the true population.



*Example of SMOTE creating synthetic samples*

Wrap Up

We wish you the best of luck with this case study. If you have any questions, please do not hesitate to reach out.