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.

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.

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.
- Are there any ways to derive additional variables that would improve the model prediction accuracy?
- What variables, if any, can be used to create business rules that can be used to decline customer's application before the model runs?

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.
- Any additional data challenges you may face given your choice of modeling methodology.
- The final list of variables that would go into your model after performing any variable selection technique that you deem necessary.
- How you would validate the model.

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.