## 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.

*I decide to include only those loans with status of 'Fully Paid' and 'Default', since the goal for the model is to predict the probability of default for each loan and only those 'Fully Paid' loans will be helpful for us to learn. The loans under 'Fully Paid' has no chance to be default any more while it might happen to those loan under 'Current' status.*

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.

**Missing values**:

I checked missing values for all columns, and remove those columns with >10% missing values, after that, for those columns that have <10% missing values, I imputed the missing values by using mean, for categorical variables, I use “Not Available” and leave them there. Since there are only one column, emp\_length(categorical) with missing values < 10%, I assume that the employment length is 0.

**Remove columns:**

There are columns that are either useless for my prediction or has leakage if I put them into the data set. Therefore, I removed these columns from my data set

1. there are columns that are not meaningful in the dataset, which means that it won't provide any information for the prediciton : id, add\_state, earliest\_cr\_line,last\_credit\_pull\_d
2. to avoid data leakage, columns like 'issue\_d', which means the issuance month of the loan, will leak infomation to the model ,besides that, columns like 'last\_fico\_range\_high', 'last\_fico\_range\_low' also have the same problems, therefore, I choose to remove these columns from the data set

**Feature Engineer:**

1. fico score variables: fico\_avg = (fico\_low +fico\_high) / 2
2. installment\_ratio: installment / (annual\_inc/12) , this shows how difficult for a borrower to pay back the loan
3. Ordinary Variable: emp\_length, map this categorical variable to numeric variable
4. other categorical variables: term, home\_ownership,verification\_status,purpose
5. feature scaling : standardization on numerical variables

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.

*Last\_fico\_range\_high,last\_fico\_range\_low,isse\_d*

* Are there any ways to derive additional variables that would improve the model prediction accuracy?

Fico\_avg, installment\_ratio

* What variables, if any, can be used to create business rules that can be used to decline customer’s application before the model runs?

dti,

loan\_amnt,

annual\_inc,

fico\_score,

installment\_ratio(loan\_amnt/(annual\_inc/12))

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.

*The problem is a binary classification so I used Logistic regression with GridSearch technique to figure out the best model, the output is a score representing the chance for each borrower to default*

* Any additional data challenges you may face given your choice of modeling methodology.

*Since the data is imbalance, I use SMOTE(synthetic Minority Oversampling Technique) to handle the imbalance data set, and reach a weight for 50% for each type of loan.*

* The final list of variables that would go into your model after performing any variable selection technique that you deem necessary.

['loan\_amnt',

'emp\_length',

'annual\_inc',

'dti',

'acc\_now\_delinq',

'delinq\_amnt',

'delinq\_2yrs',

'inq\_last\_6mths',

'fico\_avg',

'installment\_ratio',

'term\_60 months',

'home\_ownership\_MORTGAGE',

'verification\_status\_Not Verified',

'verification\_status\_Verified',

'purpose\_car',

'purpose\_debt\_consolidation',

'purpose\_home\_improvement',

'purpose\_house',

'purpose\_major\_purchase',

'purpose\_medical',

'purpose\_moving',

'purpose\_other',

'purpose\_renewable\_energy',

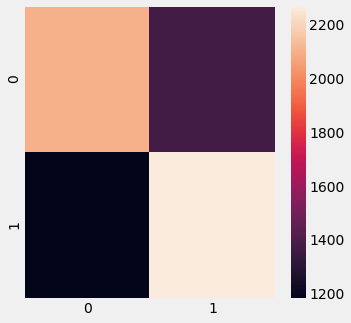
'purpose\_small\_business',

'purpose\_vacation',

'purpose\_wedding']

* How you would validate the model.

*I used ROC , AUC Score to measure the performance of the model, in addition to a confusion matrix to measure the accuracy of the model. The following chart represent the confusion matrix by my final model , coming out with a gridsearch on logistic regression model. Also, before gridsearch, I also measure the roc, auc score, which is %56 while after grid search techniques and SMOT, I got a score of %63 .*



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.