BX Data Science – Take-Home Assessment

Download Data

We will be exploring Lending Club's loan origination data from 2007-2015. Please download the dataset (loan.csv) and associated dictionary (LCDataDictionary.xlsx) from **here**.

Part 1: Data Exploration and Evaluation

For this project, please use the following columns:

```
    'loan_amnt', 'funded_amnt', 'term', 'int_rate', 'grade', 'annual_inc', 'issue_d',
    'dti', 'revol_bal', 'total_pymnt', 'loan_status'
```

Load the data, select the relevant columns, and perform any necessary cleaning and aggregations to explore and better understand the dataset. Describe any assumptions you made to handle null variables and outliers. Describe the distributions of the features. Include two data visualizations and two summary statistics to support these findings.

Part 2: Business Analysis

We are interested in evaluating whether the 36 month term loans would make for a good investment. Please investigate the following. Assume a 36 month investment period for each loan, and exclude loans with less than 36 months of data available.

1) What percentage of loans has been fully paid?
2) When bucketed by year of origination and grade, which cohort has the highest rate of defaults? Here you may assume that any loan which was not fully paid had "defaulted".
3) When bucketed by year of origination and grade, what annualized rate of return have these loans generated on average?
   For simplicity, use the following approximation:
   ```
   Annualized rate of return = (total_pymnt  / funded_amnt) ^ (1/3) - 1
   ```

Part 3: Modeling

Please build a logistic regression model to predict loan defaults (as defined in question 2 above) that could help us avoid investing in such loans. Assume that (i) you are given the ability to invest in each loan independently; (ii) you invest immediately following loan origination and hold to maturity (36 months); and (iii) all loan fields that would be known upon origination are made available to you.

Was the model effective? Explain how you validated your model and describe how you measure the performance of the model.

Submit all code related to processing, visualizing and modeling via Github. Please make sure to remove all mention of Blackstone from files you upload. You can submit either a Jupyter notebook or python files with all necessary outputs (model outputs and png files of plots). Include a documentation file with your responses and explanations for your methodology.