**E04 ML Assignment**

The data set contains complete loan data for all loans issued through the 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others. The file is a matrix of about 890 thousand observations and 75 variables. A data dictionary is provided in a separate file. Please see the Data Set Section in this assignment for specifics to be used for the data set.

**Load the data into your local laptop | Cloud account. This must be presented in live code runs. No power point or offline approach to be considered.**

1. You are required to show your capability by the following
    a. Articulate in a structured manner the approach taken
    b. Must explain this clearly (use whiteboard model | simple text pad based notes)
2. Demonstrate the understanding of the provided data set aspects
    a. Feature engineering
    b. Data Model for building the Classification (to Approve the loan request)
3. Model must be built on the data (Loan Data 2007-2011 | Declined 2007-12)) for the loan approved and declined loans
4. The model must be trained and tested with a 10-fold CV
    a. Share and explain the model accuracy
5. Once the model has been trained it must be tested against the rejected cases of 2018 Q1 to Q4
    a. Share and explain the model accuracy

**Data Set -** https://www.lendingclub.com/info/download-data.action

| Data Set | Model Build Time Frame | Model Test |
|---|---|---|
| LOAN DATA | 2007-2011 | 2008 Q1 to Q4 |
| DECLINED LOAN DATA | 2007-12 | |

DATA DICTIONARY - Link