**Decision Tree & Random Forest Problem**

For this project, we will be exploring publicly available data from [LendingClub.com](http://localhost:8888/notebooks/15-Decision-Trees-and-Random-Forests/www.lendingclub.com). Lending Club connects people who need money (borrowers) with people who have money (investors). Hopefully, as an investor, you would want to invest in people who showed a profile of having a high probability of paying you back. We will try to create a model that will help predict this.

The lending club had a [very interesting year in 2016](https://en.wikipedia.org/wiki/Lending_Club#2016), so let's check out some of their data and keep the context in mind. This data is from before they even went public.

We will use lending data from 2007-2010 and be trying to classify and predict whether or not the borrower paid back their loan in full. You can download the data from [here](https://www.lendingclub.com/info/download-data.action) or just use the CSV already provided. It's recommended you use the CSV provided as it has been cleaned of NA values.

Here are what the columns represent:

* credit policy: 1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.
* purpose: The purpose of the loan (takes values "credit\_card", "debt\_consolidation", "educational", "major\_purchase", "small\_business", and "all\_other").
* int.rate: The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.
* installment: The monthly installments owed by the borrower if the loan is funded.
* log.annual.inc: The natural log of the self-reported annual income of the borrower.
* dti: The debt-to-income ratio of the borrower (amount of debt divided by annual income).
* fico: The FICO credit score of the borrower.
* days.with.cr.line: The number of days the borrower has had a credit line.
* revol.bal: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
* revol.util: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
* inq.last.6mths: The borrower's number of inquiries by creditors in the last 6 months.
* delinq.2yrs: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
* pub.rec: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

**Here take the dataset:** [**Decision Tree & Random Forest Problem: Data Set**](https://docs.google.com/spreadsheets/d/1YKLH-mnQaKTBBrruQD1WcYLXQPiqPsvCEmkI9ciA3Pc/edit#gid=2028481626)

**Steps that you can follow:**

1. Read the dataset as a pandas dataset
2. Do some primary analysis on that by using head, info, describe etc.
3. Plot the graphs.
4. Start with scikit learn and separate X and y.
5. Train the model with decision tree and run test data
6. Use classification\_report & confusion\_matrix for evaluation
7. Train the model with random forest and run test data
8. Use classification\_report & confusion\_matrix for evaluation