Lab exercise in Python: More Robust Predictions w/ Decision Trees

Objectives

At the end of this exercise, you should be able to pull together everything we've learned to build a more robust decision tree classifier for our lending club data:

- Standardize, discretize, and transform features
- Perform basic principal component analysis
- Split data into training, test, and validation sets
- Build a simple decision tree classifier
- Examine metrics for measuring model performance
- Search the parameter space for the "best" parameters

Material

For this lab, we will continue to use Lending Club data. Lending Club is "the world's largest online marketplace connecting borrowers and investors." It is a peer-to-peer lending network that open sources some of its loan data to the community.

For this lesson, you'll need the *loan.csv* file that contains historical data on loans organized by Lending Club between 2007 and 2011. The file, along with the relevant data dictionary, can be downloaded from D2L under Week 3 documents.

In this exercise, you'll learn to explore and preprocess data in Python. The file contains numerous features, including:

- **dti** - A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
- **funded_amnt** - The total amount committed to that loan at that point in time.
- **installment -** The monthly payment owed by the borrower if the loan originates.
- **Int_rate** - The interest rate on the loan.
- **term** - The number of payment on the loan.
- **purpose** - A category provided by the borrower for the loan request.
- **grade –** A Lending Club assigned loan grade.
- **annual_inc -** The self-reported annual income provided by the borrower during registration.
- **loan_status** - The current status of the loan