Homework 2 – Machine Learning Pipeline
Due April 17, 2018

Goal:
The goal of this assignment is to build a simple, modular, extensible, machine learning pipeline in Python. The pipeline should have functions that can do the following tasks:

1. Read/Load Data
2. Explore Data
3. Pre-Process and Clean Data
4. Generate Features/Predictors
5. Build Machine Learning Classifier
6. Evaluate Classifier

As mentioned in class, the objective here is to build something simple, that is correct, can be extended later, modular, but not necessarily providing excellent results. We'll improve it over the rest of the quarter.

Data:
The data set below is a modified version of data
from https://www.kaggle.com/c/GiveMeSomeCredit

- Data Dictionary (Don't get used to it – this is a rare occurrence)
- Data File

Problem:
The task here is to predict who will experience financial distress in the next two years. The outcome variable (label) in the data is SeriousDlqin2yrs. We have access to other information about this person (as described in the data dictionary). Your assignment is to take this data and build a machine learning pipeline that trains *one* machine learning model on the data.

Don't worry too much about the model being good at this point. The primary goal is to build a skeleton code pipeline that has the components described above.

There is useful code in Jupyter notebooks in this
github repo: https://github.com/yhat/DataGotham2013/

The purpose of this homework is to start building your ML pipeline. Don't worry too much about solving the specific problem well. Focus on the overall structure, code modularity and extensibility, and getting familiar with sklearn.

What components should your pipeline have??

1. Read Data: For this assignment, assume input is CSV and write a function that can read a csv into python. It's ok to use an existing function that already exists in python or pandas.
2. Explore Data: You can use the code you wrote for assignment 1 here to generate distributions of variables, correlations between them, find outliers, and data summaries.
3. Pre-Process Data: For this assignment, you can limit this to filling in missing values for the variables that have missing values. You can use any simple method to do it (use mean to fill in missing values).
4. Generate Features/Predictors: For this assignment, you should write one function that can discretize a continuous variable and one function that can take a categorical variable and create binary/dummy variables from it. Apply them to at least one variable each in this data.
5. Build Classifier: For this assignment, select any classifier you feel comfortable with (Logistic Regression for example or Decision Trees)
6. Evaluate Classifier: you can use any metric you choose for this assignment (accuracy is the easiest one). Feel free to evaluate it on the same data you built the model on (this is not a good idea in general but for this assignment, it is fine). We haven't covered models and evaluation yet, so don't worry about creating validation sets or cross-validation.

What to submit: You should submit:

1. **link** to the code (on github) for the ML Pipeline with the components listed above in your github repository
2. a **writeup** describing what you did and results of running the code on the following problem

Some more Tips:

What I'm looking for is one (or more) python file(s) that contains all the functions of your pipeline for reading data, processing data, generating features, building models, etc. A second file for using that pipeline. You will import that pipeline-library file and use it to solve the problem you're working on for this homework.