**Assignment 1:** *Intro to ML classifiers using scikit and Jupyter Notebook*

**GOAL:** Build a predictive machine learning classifier that will use historic data of loan decisions (ground truth) to make recommended decisions for new loan applications. Along the way, you’ll get comfortable with using Jupyter Notebooks to document your work, using scikit-learn to train and test basic ML models, and start thinking about the strengths and weaknesses of models like this.

In this assignment we’ll get started with a classic ML problem: *determining whether customers will get approved or rejected for loans based on a variety of factors*.

**Background:** In the US (and a number of other countries), it is common for people who want to buy a home to apply for a loan (called a “mortgage”) from a bank or other lender. In order for the bank to decide whether to give them this loan, the bank will look at a number of factors to determine whether they are likely to be able to repay the loan because the bank does not want to lend money to somebody who would not be able to pay it back. These factors typically include how much money the person makes, how educated they are, how much money they are asking for in the loan, and a handful of other things.[[1]](#footnote-0) Banks typically have formulas in place to determine who they will approve for a given loan and who they will deny, but there is also frequently some human judgment involved in the process.

If the person is approved for the loan, they will use the loan to purchase the home and will pay it back in small chunks every month for a set period of time (usually 15 or 30 years). If they are denied, they may have to find a less expensive home to buy or address some of the factors that led them to be denied (e.g., they might have to wait until they make more money).

In this task, we are given a set of data about whether customers were approved or denied a loan based on a number of factors. In this assignment, your job will be to build a model that predicts whether a person will be approved or denied based on these same factors.

Setup Logistics:

Checklist:

1. Install Python 3.6 or greater (ideally w/ Anaconda)
2. Install Jupyter Notebook
3. Install packages

This will be written assuming use of Python 3.6 or greater. We **strongly** recommend using [Anaconda](https://www.anaconda.com/) for this. You’ll also need to install [Jupyter Notebook](http://jupyter.org/install) (which comes with Anaconda) and [scikit-learn](http://scikit-learn.org/stable/install.html), [matplotlib](https://matplotlib.org/), [pandas](https://pandas.pydata.org/), and [numpy](http://www.numpy.org/). We’re not going to lead you through the installation of these, but there are many YouTube tutorials that go through the process step-by-step.

If you’re new to Jupyter Notebook, there’s a decent (though slightly dated) walkthrough [here](https://www.youtube.com/watch?v=HW29067qVWk). The purpose of using Jupyter Notebook here is both to document all of your steps and to allow you to mix code, output, and commentary in the same space.

Assignment Pieces

There are five files in the [A1 folder](https://drive.google.com/drive/folders/1qeR8Rg2Be0Z2sUFSD5tbwqJyrK7wwoKm?usp=sharing)[[2]](#footnote-1):

* [**loantrainingdata.csv**](https://drive.google.com/file/d/11U5CMia9d7m3VJLh6swt3O4ag2IR7UtD/view?usp=sharing) - the data you will train your model on. This is data showing a variety of attributes about people who applied for loans and whether they were approved for or denied the loan.
* [**loantestset.csv**](https://drive.google.com/file/d/1pHst5YyR3B2kIoNiIuv8WXmwPQPQ52te/view?usp=sharing) - the data you’ll use to test your model, to see how good your model is at predicting who got a loan.
* [**predictiondata.csv**](https://drive.google.com/file/d/1_rf2nTJyLtulCeZKFwBWLc40x2VvEadu/view?usp=sharing) - another dataset, except this time you don’t have the answers to see how well you did (and neither do we). You’ll make some predictions without knowing the “right” answers, and then you’ll talk about what this means.
* [**variables.txt**](https://drive.google.com/file/d/1ebRiHA8UuQeCOS300DxWUeFSnGWlDYm8/view?usp=sharing) - a list of the variables in the files and explanations of what they mean
* [**HAIIF18\_a1.ipynb**](https://drive.google.com/file/d/15wP7yyUQDoU396r1qx8NwNqO1zC8WOOn/view?usp=sharing)- a starter Jupyter Notebook that will list the rest of this assignment.

Making a model

The remainder of this assignment will be in the **HAIIF18\_a1.ipynb** notebook. Save a copy of this file for yourself, rename it, and fill in the missing pieces and answer the questions as directed.

Please download your final notebook as HTML, and submit to the Canvas site with name [yourandrewid]\_haiif18a[assignmentnumber], e.g., jseering\_haiif18a1.

1. Note that interest rates are also major factors, but for this assignment we are assuming that the interest rate for all of the loans was the same, so it is not included. [↑](#footnote-ref-0)
2. This data is a modified form of the data from [here](https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/). [↑](#footnote-ref-1)