Some notes before starting:

\* Read all the way through the instructions.

\* Models must be built using Python, R, or SAS.

\* New features can be created.

\* Users cannot add or supplement with external data.

\* While simple techniques may develop adequate models, success in this exercise typically involves feature engineering and model tuning.

\* Throughout your code, please use comments to document your thought process as you move through exploratory data analysis, feature engineering, model tuning, etc.

\* Please review your submission against the submission expectations.

Step 1 - Clean and prepare your data:

There are several entries where values have been deleted to simulate dirty data. Please clean the data with whatever method(s) you believe is best/most suitable. Note that some of the missing values are truly blank (unknown answers). Success in this exercise typically involves feature engineering and avoiding data leakage.

Step 2 - Build your models:

Please use two different machine learning/statistical algorithms to develop a total of two models. Please include comments that document choices you make (such as those for feature engineering and for model tuning).

Step 3 - Generate predictions:

Create predictions on the data in test.csv using each of your trained models. The predictions should be the class probabilities for belonging to the positive class (labeled '1').

Be sure to output a prediction for each of the rows in the test dataset (10K rows). Save the results of each of your models in a separate CSV file. Title the two files 'results1.csv' and 'results2.csv'. A result file should each have a single column representing the output from one model (no header label or index column is needed).

Step 4 - Compare your modeling approaches:

Please prepare a relatively short write-up comparing the pros and cons of the two algorithms you used (PDF preferred). As part of the write-up, please identify which algorithm you think will perform the best. For the best performing model, are there choices you made in the context of the exercise that might be different in a business context? How would explain to a business partner the concept that one model is better than the other?

Step 5 - Submit your work:

Your submission should consist of all the code used for exploratory data analysis, cleaning, prepping, and modeling (text, html, or pdf preferred), the two result files (.csv format - each containing 10,000 decimal probabilities), and your write-up comparing the pros and cons of the two modeling techniques used (text, html, or pdf preferred). Note: The results files should not include the original data, only the probabilities.

Your work will be scored on techniques used (appropriateness and complexity), evaluation of the two techniques compared in the write-up, model performance on the data hold out - measured by AUC, and your overall code/comments. The threshold for passing model performance is set high, expecting that model tuning and feature engineering will be used. The best score of the two models submitted will be used.

Please do not submit the original data back to us.