Some notes before starting:

- Read all the way through the instructions.
- Models must be built using Python/r.
- No additional data may be added or used.
- 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.

Data sets:

To build the model, please use model.csv file
To validate the model, please use val.csv file

Goal: Correctly predict default candidates (1 means default, 0 means non-default) using val.csv file

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. Success in this exercise typically involves feature engineering and avoiding data leakage.

Step 2 - Build your models:
We request to make two models: Logistic Regression and any other machine learning/statistical models to correctly predict 'default' candidates.

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 val.csv using each of your trained models. The predictions should be the class probabilities for belonging to the default class (labeled '1').

Be sure to output a prediction for each of the rows in the validation dataset (val.csv). 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 modeling techniques you used (PDF preferred). Are there choices you made in the context of the exercise that might be different in a business context?

Step 5 - Submit your work:
Your submission should consist of all the code used for EDA, cleaning, prepping, and modeling (text, html, or pdf preferred), the two result files (.csv format), and your write-up comparing the pros and cons of the two modeling techniques used (text, html, or pdf preferred).

Your work will be scored on the following:

- 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
- 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.