Take-Home Technical Assessment

A British challenger bank collected data about the performance of their loans issued from 2009 to 2015. They would like to use the data to build a model to help their credit analysts assess new loan applications, i.e. predict if a loan would be repaid in full or defaulted on. The *loans.csv* dataset contains ~230k examples and 31 different variables. A description of these variables is provided in *data-dictionary.csv*.

You are required to analyse these data and prepare a short presentation (15 minutes) explaining your analysis and conclusions. This should include any quantitative (exploratory data analysis or summary statistics) or qualitative (interpretation and context) insights gained from the data. Please note this is a real dataset and so may require some cleaning.

Your analysis must include, <u>but need not be limited to</u>, the following points:

1) Develop a model (*Model-1*) that, given all available variables, predicts the probability of default at the time of application;
2) Identify which variable is the most predictive of loan performance;
3) Identify any highly correlated pair of variables;
4) Develop a second model (*Model-2*) that uses only 5 of the provided variables – explain your choice;
5) Compare *Model-1* and *Model-2* – explain your choice of metrics;

You are required to provide *all the files containing your workings* (Excel, R, Python, Stata, Matlab etc… - choose whichever you prefer) *and the presentation (ppt/pdf)*. You'll be asked to present your results during the first 15 minutes of your face-to-face interview. This will be followed by a 45-minutes technical interview.

You will be assessed based on the technical rigour and robustness of your analysis, the ability to provide new and non-trivial insights and your communication and presentation skills.