**LBG IP&I Data Science Take Home Exercise**

This is a take home data science exercise, with an ethical machine learning component.

The csv data file provided contains synthetic data representing a stratified random sample of customers who have a Lloyds Banking Group savings account, and who satisfy several other criteria: they are all aged 18 or over, none of them have retired yet, and all of them have a current account with Lloyds Banking Group that (based on separate analysis) appears to be their main current account. The data are for a single year only. A data dictionary is provided containing definitions of all columns. All monetary amounts are in GBP and are full-year amounts.

Please perform the pre-processing steps and analyses requested below using the python or R programming language (python is preferred). Return code scripts or code notebooks plus supporting writeup (your writeup can be in the code notebook) to us within one week of receiving these instructions. Keep the writeup short and concise – there are no points for high word count!

Pre-processing and analysis required:

1. Cleanse the data and prepare it to be suitable for statistical modelling and machine learning. Comment on the steps performed and explain the rationale for them.
2. Produce descriptive statistics and plots describing the individual variables and their pairwise associations with one another. Pay particular attention to the associations of income and annual savings with all other variables. Comment on the results. Is there anything particularly interesting? Is there anything that doesn’t make sense?
3. Build statistical model(s) – also known as econometric model(s) – to explain how a set of explanatory variables each affect annual net savings. Comment on the results. Which explanatory variables have robust statistically significant effects in your model(s)? Do these results make sense?
4. Build supervised machine learning model(s) to predict annual net savings based on other variables. Include appropriate safeguards against overfitting. Comment on the results. How well do the models perform? Is the performance stable? Are the models interpretable / explainable? Are these models suitable for decision-making? Is there enough signal in the data to build good models? Is there anything suspicious about the results?
5. Are there any ethical concerns with using any of the predictor variables in the dataset? Which variables, and why? Would you change the model to address these concerns? Does this change have a cost in terms of model performance? How would you balance performance against ethics? Which model would you finally decide to use to predict customer savings rates, and why? Is there anything else ethically questionable about the data?

Good luck!