Kellogg Data Scientist Technical Exercise

**Data**

For this task, you’ll have access to 1-year of simulated sales data with supplementary information. This is captured within the four tables:

* Sales.csv – product purchase data aggregated at customer, month, store, and product level
* Product.csv – product information
* Stores.csv – store descriptions
* Customer\_supplement.csv – shopper information that can be assumed to be obtained via an optional questionnaire.

In this fictitious example, you can assume that the products you represent have a positive ‘K flag’. Where this is FALSE, these are competitors.

**Task**

Your stakeholders are keen to understand which customers are not likely to purchase a ‘K’ product within the next 3-months. They would like you to be able to predict the likelihood that this occurs, given a customer has purchased a ‘K’ product within the past 9-months.

Whilst carrying this out, please consider any data quality issues and approach it with exploratory data analysis like you would with any other project.

The stakeholders are also interested in developing a personalised marketing campaign for groups of customers. *Please advise on a potential set of actions from scoping through to implementation.*

Feel free to make assumptions about the data or task but be clear about why and what you’ve done as a result.

You will be presenting back on your findings to a panel of Data Scientists, in which you will have 30 minutes to present with a further 30 minutes for follow-up questions.

**Expectations**

Feel free to use the programming language that you’re most comfortable with.

What we’re looking for:

* Ability to wrangle data
* Knowledge of appropriate data science techniques
* Curiosity and story-telling ability

We’re not expecting a refined solution (i.e. no model tuning) but would like to gain insight into your technical abilities and thought process. It is sufficient to note the actions you would take if given more time on this task.