**Data Scientist**

**Applicant Assessment**

***A note to candidates:*** *This assessment was designed to assess two crucial practices; creating solutions, and reviewing solutions so as to improve or extend them. This assessment is intended to take approximately two hours to complete!*

**Introduction**

You are collaborating a fellow Data Scientist on a business-critical R&D project. The project is concerned with modelling fraudulent behaviour so that such behaviour can be detected. The goals for the model are to predict fraud, and to prevent insult to legitimate customers.

On your first day our colleague provides you with the following materials that they prepared ahead of time:

* A dataset to use in developing the model (attached)
* A code solution developed within ipython/jupyter (attached)
* A brief overview on the objectives of the current iteration (under the heading **Project** **Objectives** below).

After some Q&A with your colleague, you get some added details about the materials your colleague provided. This detail can be found below, in the section **Notes on the Materials Provided**.

**Notes on the Materials Provided**

The dataset provided (*traffic.csv*) is synthetic, labelled transaction data for a set of accounts. The data contains an account identifier (accountid\_hashed).

This transactional data has been extracted from our team’s Feature Store; which means that this data was initially prepared into features by our team’s feature creation pipeline. Your colleague has provided a data dictionary with information on the features available in the traffic dataset.

The code solution provided is written using Python, in an iPython Jupyter environment. The code is supported by inline comments and markdown documentation. .py files are provided as an alternative – these contain inline comments but less markdown documentation.

Your colleague attempted a modelling solution in sklearn, and a separate solution in Keras.

**Project Objectives**

As above, the aims of this project are to prevent fraud, and to prevent insult to customers. Preventing fraud is about correctly identifying cases of fraud, and preventing insult is about not incorrectly identifying fraud.

The business has a target for fraud prevention, your colleagues in the Marketing team have extracted a target from your client of [x]% correct identification.

The client in question is extremely keen to reduce insult to a value as close to 0 as possible.

A good solution is able to perform both tasks effectively or find an appropriate tradeoff between both.

**Assessment Questions**

1. Review the code solution provided by your colleague. What feedback would you give to your colleague to help them improve any future modelling work?

In your answer, consider whether to reference elements of your colleague’s workflow, technique selection choices made and possible alternatives, modelling approach, evaluation approach, code quality and style.

1. Using the data provided, construct an MVP code solution. Prioritize the most important solution elements. Explain the reasoning behind decisions to implement or not implement.