**Advanced Analytics Centre – Machine Learning Engineer Case Study**

The purpose of this task is to test:

* Your knowledge of machine learning principles
* Your knowledge of coding standards and the layout of production code

Present your approach to a combination of Technical and Business people for around 20-30 minutes covering:

* + Brief Summary of the Problem
  + The step-by-step approach to how you progressed, any assumptions made
  + Tools and infrastructure used and how you set it up
  + Any challenges you encountered
  + Provide the code that you wrote to answer the problem
  + Any other observations about the model or model code

You will be given up to 5 days to prepare your solution.

Problem Statement

The main cause of road casualties is related to driver behaviour. The identification of unsafe driving is an important step in order to take actions to the number of road casualties.

The Road Safety Authority has made it mandatory for drivers to have a device installed in their vehicles that displays the driver’s ‘aggressiveness score’ real-time during trips.

A graduate data scientist has built a predictive machine learning model where the output is an aggressiveness score based on various metrics, including fuel consumption, rpm, speed, etc.

The next step is to productise\* the model so that car manufacturers can integrate it with the standard car software, and you have been assigned to develop this solution.

You are required to:

* + Make the code production-ready
  + If relevant, pick up on any (methodical) flaws
  + Make the code meet coding standards
  + Explain ongoing monitoring and maintainability

*Productionisation can be thought of as “A systematic process of taking a model, integrating it with business processes in an automated way, making it available in a reproducible, reliable and scalable manner. A productionised model is well integrated into a software system, is well tested, is monitored and has ongoing support.”*

Data Description

The data folder contains all the data that was extracted from source systems. **Only use the datasets that are directly referenced in the model code.**

1. Drive Data (connected car data) – Data coming from the car-mounted devices, which provides you with the car statistics every second. This information will include – Speed, acceleration, engine temperature and other car statistics.

2. Trip – Parameters associated with location of car such as latitude, longitude, altitude and other similar parameters

3. drive\_features.csv – The drive features have already been produced for you and are available as a CSV file.

Data Location

<https://drive.google.com/open?id=1wOQV7yNn0F240UJDY1oEEs0nmqQTPUnF>

Code Location

<https://bitbucket.org/wesaac/de_case_study/src/master/models/case_study.R>

Some additional (optional) Reading:

Source: [Hidden Technical Debt in Machine Learning Systems](https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf)

